Identificabilidad de modelos matemáticos en procesos biotecnológicos

Julio César Sánchez Rendón

Universidad Nacional de Colombia
Facultad de Ingeniería y Arquitectura, Departamento de Ingeniería Química
Manizales, Colombia
2021

Identificabilidad de modelos matemáticos en procesos biotecnológicos

Julio César Sánchez Rendón

Tesis presentada como requisito parcial para optar al título de:
Magister en Ingeniería - Ingeniería Química

Director:
Ph.D. Oscar Andrés Prado Rubio

Codirector:
Ph.D. Luis Gerónimo Matallana Pérez

Línea de Investigación:
Modelamiento matemático y simulación
Grupo de Investigación:
Grupo de Investigación en Aplicación de Nuevas Tecnologías - GIANT

Universidad Nacional de Colombia
Facultad de Ingeniería y Arquitectura, Departamento de Ingeniería Química
Manizales, Colombia
2021

Prefacio

Esta disertación fue escrita como un requisito parcial para el título de Magister en ingeniería (MSc) - Ingeniería Química. Este proyecto fue llevado a cabo desde agosto de 2017 hasta abril de 2021. Esta tésis fue elaborada en el Grupo de Investigación en Aplicación de Nuevas Tecnologías (GIANT) en el Departamento de Ingeniería Química de la Universidad Nacional de Colombia, sede Manizales.

Primero, quisiera agradecer a mi director el Profesor Asociado Oscar Andrés Prado Rubio y co-director el Docente de Departamento Luis Gerónimo Matallana Pérez por su acompañamiento, motivación y paciencia a lo largo de este trabajo de maestría. Así mismo, le extiendo estos agradecimientos al Profesor Ricardo Morales Rodriguez y a la Universidad de Guanajuato por haberme acogido en una estancia internacional de investigación. De manera especial, quiero agradecer al profesor Oscar por su dedicación y apoyo más allá de la labor docente, por creer en mis capacidades y mi talento y haberme motivado constantemente a dar lo mejor de mí mismo.

Agradezco también a mi familia por su amor y apoyo incondicional y dedico este trabajo a mi madre y mi padre, que los logros que hemos cosechado les llenen de orgullo. Agradezco a mis compañeros Jason, Laura, Luis, Carlos, Oscar y Miguel por las risas, las discusiones y los buenos momentos de esta travesía. Agradezco a Daniel, Ricardo, Jhony y Camilo por nuestra hermandad, las risas y las peleas, por haberme acompañado en todo este camino de más de 10 años. Agradezco a Mayra y a Ebtisam por haberme escuchado, aconsejado, acompañado, por haber estado a mi lado en el momento más oscuro y ayudarme a ver que la vida continua.

Finalmente, agradezco a Marcela por haberme acompañado a lo largo de mi camino en la academia y la vida, por las risas, las lágrimas y las discusiones, por el amor incondicional y por haberme enseñado la lección más valiosa de mi vida. Espero que algún día nuestros caminos se vuelvan a cruzar y la historia continúe.

Julio César Sánchez Rendón
Manizales, Octubre, 2021

Resumen

Identificabilidad de modelos matemáticos en procesos biotecnológicos

La biotecnología puede ser una opción viable para mitigar o resolver problemáticas generadas por la acción antrópica, mediante la generación de alternativas más eficientes y ambientalmente amigables a los procesos industriales convencionales. Desde la perspectiva de ingeniería de sistemas de proceso, los sistemas biológicos pueden describirse como modelos matemáticos útiles para el diseño, control y optimización de procesos biotecnológicos. Debido a esto, se genera la necesidad de obtener modelos matemáticos con capacidad descriptiva y predictiva. Sin embargo, los modelos de bioprocesos presentan un reto desde el punto matemático por diferentes circunstancias como su naturaleza no lineal, incertidumbre en mediciones experimentales, sobreparametrización, entre otros.

Esfuerzos previos para construcción de una metodología sistemática han sido propuestos, como el marco teórico desarrollado por Cameron y Hangos en su libro "Process Modelling and Model Analysis". En relación a lo anterior, esta tesis tiene como propósito la consolidación de una metodología extendida para generar modelos matemáticos de bioprocesos con interpretabilidad y capacidad tanto descriptiva como predictiva. El caso de estudio seleccionado corresponde a bioproducción de xilitol, que considera características inherentes a los sistemas biológicos como inhibición y transporte de metabolitos.

Como punto de partida se analizaron los datos experimentales. En ese sentido, se investigaron métodos para detección y limpieza de puntos atípicos junto con un método de ajuste polinomial para eliminación de ruido aleatorio presente en las mediciones experimentales. Como resultado, se obtuvo una metodología para limpieza y suavizados de datos experimentales, así como un índice de calidad para los datos suavizados.

Subsecuentemente, para estimar los parámetros de un modelo matemático es necesario que dichos parámetros sean *identificables*. Esto implica que su valor pueda ser conocido a partir de mediciones experimentales. Sin embargo, la identificabilidad de un parámetro depende tanto de datos experimentales como de la estructura misma del modelo matemático. Para determinar las propiedades identificabilidad del modelo matemático de bioproducción de xi-

litol, se recopilaron métodos y software de identificabilidad estructural y se realizó el análisis sobre el caso de estudio considerando los datos experimentales disponibles y sus características. Se encontró que el modelo es estructuralmente localmente identificable con condiciones iniciales no nulas de glucosa y xilosa, lo que añade importancia adicional al diseño experimental enfocado a estimación de parámetros.

Seguidamente, se investigó la identificabilidad práctica de parámetros. Con este propósito, se estudiaron la función objetivo y el algoritmo de optimización los cuales permiten *ajustar* el comportamiento del modelo matemático a través de la comparación de sus predicciones con datos experimentales. Un tipo particular de algoritmos de optimización conocido como metaheurísticas (de naturaleza estocástica) han ganado popularidad por su facilidad de uso y eficiencia. Sin embargo, las metaheurísticas poseen parámetros internos que controlan su comportamiento en el espacio de búsqueda. En relación con lo anterior, se analizó la influencia de la sintonización del optimizador y el tipo de normalización de la función objetivo en la precisión y reproducibilidad de la solución del problema de optimización para estimación de parámetros del modelo de bioproducción de xilitol. A través de la generación de una herramienta híbrida para sintonización de metaheurísticas en Matlab® por interconexión con R®, se demostró que la sintonización mejora significativamente la reproducibilidad y precisión de la solución del problema de optimización. Además, se encontró que la combinación sintonizada de optimizador enjambre de partículas (PSO) con factor de normalización por media de variable experimental presenta el mejor desempeño para el modelo de caso de estudio.

Finalmente, se estudió la influencia de los diferentes elementos del problema de optimización en el valor de los parámetros del modelo matemático analizado. Así mismo, la calidad de la predicción del modelo de caso de estudio se examinó con intervalos de confianza, indicadores de ajuste e indices de sensibilidad. Específicamente, se determinó que el optimizador enjambre de partículas sintonizado pudo encontrar satisfactoriamente el mínimo global de la función objetivo y el pretratamiento de datos experimentales reduce la incertidumbre en el valor de los parámetros. Además, se demostró que aunque un conjunto de datos presente identificabilidad estructural no necesariamente también brindará identificabilidad práctica. En cuanto a la validación del modelo, se encontró alta interacción entre parámetros debido a carencia de datos experimentales con concentraciones iniciales de glucosa y xilosa, se analizó la sensibilidad de indicadores de ajuste, se determinó la incertidumbre en la respuesta del modelo y se calculó la sensibilidad del modelo a incertidumbre en los parámetros, siendo mejor el método de coeficientes de regresión estandarizados (SRC) sobre los índices de Sobol. A partir de esta información se estableció una metodología de estimación de parámetros y validación de modelos.

La metodología propuesta como resultado de esta investigación apunta a la generación de modelos matemáticos de procesos biotecnológicos robustos, con interpretabilidad en sus

parámetros y con capacidad tanto descriptiva como predictiva, útiles para el diseño, control y optimización de bioprocesos. Esta metodología aborda e incluye aspectos que hasta el momento presente no han sido extendidos o estudiados ampliamente en el modelamiento de procesos biotecnológicos, entre los que se encuentran: limpieza de datos experimentales que reduce incertidumbre en el valor de los parámetros, identificabilidad estructural que da certeza teórica en la selección de datos experimentales, sintonización de algoritmos de optimización que mejora la precisión y reproducibilidad de la estimación de parámetros y finalmente, identificabilidad práctica divida en calidad descriptiva y predictiva que cuantifica la calidad del modelo matemático en cuanto a su calibración y uso en aplicaciones prácticas.

Palabras clave: estimación de parámetros; modelo matemático; validación de modelos; xilitol; metodología.

Abstract

Identifiability of mathematical models in biotechnological bioprocesses

Biotechnology can be a viable option to mitigate or solve problems generated by anthropogenic action, by generating more efficient and environmentally friendly alternatives to conventional industrial processes. From the perspective of process systems engineering, biological systems can be described as mathematical models useful for the design, control and optimization of biotechnological processes. Because of this, there is a need to obtain mathematical models with descriptive and predictive capabilities. However, bioprocess models present a challenge from the mathematical point of view due to different circumstances such as their nonlinear nature, uncertainty in experimental measurements, overparameterization, among others.

Previous efforts to build a systematic methodology have been proposed, such as the theoretical framework developed by Cameron and Hangos in their book "Process Modelling and Model Analysis". In relation to the above, the purpose of this thesis is the consolidation of an extended methodology to generate mathematical models of bioprocesses with interpretability and both descriptive and predictive capacity. The selected case study corresponds to xylitol bioproduction, which considers inherent characteristics of biological systems such as inhibition and metabolite transport.

As a starting point, the experimental data were analyzed. In that sense, methods for outlier detection and cleansing were investigated together with a polynomial adjustment method for elimination of random noise present in the experimental measurements. As a result, a methodology for cleansing and smoothing of experimental data was obtained, as well as a quality index for the smoothed data.

Subsequently, to estimate the parameters of a mathematical model it is necessary that these parameters are *identifiable*. This implies that their value can be known from experimental measurements. However, the identifiability of a parameter depends both on experimental data and on the structure of the mathematical model itself. To determine the identifiability properties of the mathematical model of xylitol bioproduction, structural identifiability methods and software were collected and the analysis was performed on the case study considering the available experimental data and its characteristics. It was found that the model is structurally locally identifiable with non-zero initial conditions of glucose and xylose, which adds additional importance to the experimental design focused on parameter estimation.

Next, the practical identifiability of parameters was investigated. For this purpose, the objective function and the optimization algorithm were studied, which allow the behavior of the mathematical model to be adjusted by comparing its predictions with experimental data. A particular type of optimization algorithms known as metaheuristics (stochastic in nature) have gained popularity for their ease of use and efficiency. However, metaheuristics have internal parameters that control their behavior in the search space. In relation to the above, the influence of the optimizer tuning and the type of normalization of the objective function on the accuracy and reproducibility of the solution of the optimization problem for parameter estimation of the xylitol bioproduction model was analyzed. Through the generation of a hybrid tool for metaheuristic tuning in Matlab® by interconnection with R®, it was shown that tuning significantly improves the reproducibility and accuracy of the solution of the optimization problem. Furthermore, it was found that the tuned combination of Particle Swarm Optimization algorithm (PSO) with normalization factor by experimental variable mean presents the best performance for the case study's model.

Finally, the influence of the different elements of the optimization problem on the value of the parameters of the mathematical model analyzed was studied. Likewise, the quality of the prediction of the case study's model was examined with confidence intervals, fit indicators and sensitivity indixes. Specifically, it was determined that the tuned particle swarm optimizer was able to successfully find the global minimum of the objective function and the pretreatment of experimental data reduces the uncertainty in the value of the parameters. In addition, it was shown that even if a data set exhibits structural identifiability it will not necessarily also provide practical identifiability. Regarding the validation of the model, high

interaction between parameters was found due to the lack of experimental data with initial concentrations of glucose and xylose, the sensitivity of adjustment indicators was analyzed, the uncertainty in the model response was determined and the sensitivity of the model to uncertainty in the parameters was calculated, being better the method of standardized regression coefficients (SRC) on Sobol's indexes. Based on this information, a methodology for parameter estimation and model validation was established.

The methodology proposed as a result of this research aims at the generation of robust mathematical models of biotechnological processes, with interpretability in their parameters and with both descriptive and predictive capacity, useful for the design, control and optimization of bioprocesses. This methodology addresses and includes aspects that so far have not been extended or studied extensively in the modeling of biotechnological processes, among which are: cleansing of experimental data that reduces uncertainty in the value of the parameters, structural identifiability that gives theoretical certainty in the selection of experimental data, tuning of optimization algorithms that improves the accuracy and reproducibility of parameter estimation and finally, practical identifiability divided into descriptive and predictive quality that quantifies the quality of the mathematical model in terms of its calibration and use in practical applications.

Keywords: parameter estimation; mathematical model; model validation; xylitol; methodology.

Contenido

Capítulo 1

Introducción

La humanidad en su afán de satisfacer sus necesidades ha provocado y acentuado problemáticas ambientales, económicas y sociales. Dentro de estas se encuentran principalmente cuatro "mega-problemas" que comprenden: destrucción de la naturaleza, cambio climático, violencia política y pobreza. La destrucción de la naturaleza abarca la contaminación de los ecosistemas naturales por desechos industriales y actividad antropogénica (Moreira et al., 2016), deforestación a gran escala (Grieg-Gran et al., 2006), sobreexplotación de especies comerciales (Rosser & Mainka, 2002), perdida de la biodiversidad (Spatz et al., 2017) y disminución de servicios ecosistémicos (Tajam et al., 2017). Esta problemática se conecta con el cambio climático, principalmente relacionada con el aumento de los gases de efecto invernadero (vapor de agua, CO_2, NO_x y CH_4) en la atmósfera, desechos comunes de las actividades industriales y sistemas de transporte. Entre ellos es de especial importancia el CO_2 generado principalmente por el uso y la extracción de combustibles fósiles.

Otras problemáticas asociadas al cambio climático corresponden a la escasez de agua potable y el abastecimiento alimentario, los cuales se verán acentuados por el aumento poblacional acelerado que se predice llegará los 9500 millones de habitantes en el planeta para el año 2050 (Gerland et al., 2014). Así mismo, se verá influenciada la pobreza y la violencia política al aumentar la brecha en la distribución de la riqueza, junto con carencia aún más severa de los recursos naturales disponibles (Szombatfalvy, 2010).

Para enfrentar los efectos negativos que estas problemáticas ya están ocasionando y cuyos efectos a futuro serán devastadores, diversos actores a nivel internacional han establecido agendas que buscan mitigar y contener dichos efectos. Entre ellas se encuentra la Comisión Europea con sus estrategias "Europa 2020" y "Europa 2050" (Comisión Europea, 2010, 2018).

La estrategia "Europa 2020" busca el aumento de la cobertura de empleo formal, una reducción de al menos 20 millones de personas en condiciones de pobreza, aumento en la

cobertura de educación superior, aumento en el presupuesto de investigación y desarrollo, y reducción de por lo menos un 20 % en la emisión de gases de efecto invernadero junto con un suplemento del 20 % de la energía con fuentes renovables. Por otr aparte, la estrategia "Europa 2050" fija las metas para una transición entre el uso de combustibles fósiles hacia fuentes renovables para el año 2050. En esta estrategia se incluye la maximización de la eficiencia energética, expansión de energías renovables y uso de electricidad, implementación de movilidad limpia, segura y conectada, transición hacia una economía circular competitiva, expansión de la bio-economía y la creación de sumideros esenciales de carbono.

La biotecnología se ha consolidado como una alternativa atractiva ante la influencia adversa los problemas ya mencionados, al abarcar un amplio número de campos de acción entre los que se encuentran desarrollo de fármacos, mejora genética de plantas de cultivo y ganado, producción de energías renovables, descontaminación, producción de sustancias químicas como biosurfactantes, biopolímeros, biocatalizadores, entre otros (Sai et al., 2011; Sarmah et al., 2018; Rana et al., 2019). De igual forma, se han encontrado algunas moléculas con base biotecnológica que poseen la capacidad de reemplazar productos químicos de primera necesidad, también conocidos como *"commodity chemicals"*, obtenidos del petróleo y que se caracterizan por su alto nivel de producción y bajo costo.

Entre estas moléculas de base biotecnológica se tienen ácidos orgánicos como el láctico, succínico, fumárico y málico, los alcoholes glicerol, sorbitol y xilitol, el éster 3-hidroxibutirolactona, biocombustibles, entre otros. Cada una de estas sustancias químicas pueden ser transformadas a diferentes compuestos con diversas utilidades industriales, ganándose la denominación de "compuestos plataforma" (Werpy & Petersen, 2004; Shylesh et al., 2017). Dado que el mundo se aproxima al pico de producción de petróleo en donde se presenta agotamiento en las reservas junto a un aumento sostenido en su precio (Energy Insights - McKinsey, 2019), la posibilidad de generar *commodity chemicals* a partir de *building blocks* ambientalmente amigables y de bajo costo demuestra la conveniencia de los procesos biotecnológicos.

Actualmente, existe la necesidad de desarrollar procesos biotecnológicos sostenibles, es decir, procesos más eficientes, económicamente viables y ambientalmente amigables. La sostenibilidad puede lograrse a través de optimización en el diseño y operación de los procesos biotecnológicos, lo que requiere de la implementación de modelos matemáticos que describan de forma precisa los fenómenos involucrados en ellos (Doelle et al., 2009).

Según la filosofía de ingeniería de procesos, los modelos matemáticos hacen parte fundamental para el diseño, el control y la optimización. Como lo describen Cameron & Hangos (2001) "El modelamiento matemático une un conjunto de ecuaciones M que representan un sistema S con un propósito P. Una serie de experimentos $\mathcal{E}$ puede ser aplicados a M a fin de resolver

preguntas sobre S". En otras palabras, los modelos matemáticos permiten el entendimiento de un sistema particular y en este caso, de los procesos biotecnológicos.

Una clase particular de modelos matemáticos que tiene una amplia aplicación en el área de bioprocesos corresponde a los modelos basados en principios físicos, compuestos principalmente por ecuaciones diferenciales tanto ordinarias como parciales, así como ecuaciones algebraicas e integrales o una mezcla de todas ellas, capaces de describir el comportamiento del sistema en el transcurso de una o más dimensiones (Tasseff & Varner, 2010). Ejemplos de modelos matemáticos aplicados sistemas biotecnológicos incluyen modelamiento de sistemas fototrópicos para crecimiento de microalgas (producción de biocombustibles) (Pfaffinger et al., 2019), crecimiento en biorreactor de células de ovario de hamster chino (producción de proteínas recombinantes eucariotas) (Tang et al., 2020), catálisis enzimática con cristalización reactiva para producción y purificación de antibióticos betalactámicos (McDonald et al., 2019), producción de biogas (Neba et al., 2020), producción de biobutanol (Zhou et al., 2020), entre otros.

Debido a la necesidad de mayor capacidad predictiva de los modelos matemáticos, grandes esfuerzos se han realizado para su construcción, estimación de parámetros y validación. En particular, la estimación de parámetros requiere el uso de métodos de optimización que minimizan la desviación de la predicción del modelo respecto a los datos experimentales, mientras que la validación asegura de manera estadística la calidad predictiva del modelo (Cameron & Hangos, 2001).

Los modelos matemáticos usados en procesos biotecnológicos sufren de una alta complejidad matemática expresada en no linealidades, elevado número de ecuaciones, sobreparametrización, carencia de información sobre los fenómenos internos del sistema, problemas de bifurcación, error en las medidas experimentales, entre otros (Brun et al., 2001). Estas problemáticas dificultan tanto la estimación de parámetros como la validación de los modelos, pues el alto número de variables y ecuaciones junto con la complejidad de las mismas, deriva en un problema de optimización no convexo.

Este tipo de problemas de optimización se caracterizan por la presencia de múltiples mínimos locales y un (posible) mínimo global, por tanto, la búsqueda de los parámetros del modelo matemático requiere de algoritmos de optimización global. Existen dos grandes categorías para estos algoritmos de optimización: estocásticos que excluyen regiones de mínimos locales mediante corridas de confirmación con valores aleatorios hasta encontrar el óptimo global de la función (Banga et al., 2004), y los determinísticos que se basan en información del modelo matemático como por ejemplo, su gradiente (Cappuyns et al., 2009). Dado que los valores de los parámetros obtenidos a través de optimización poseen un grado de incertidumbre, se hace necesario el proceso de validación. Este proceso examina las capacidades descriptivas y

predictivas del modelo matemático obtenido (Englezos & Kalogerakis, 2000).

Diferentes metodologías para estimación de parámetros y validación de modelos matemáticos han sido planteadas en el transcurso del tiempo, como la propuesta por Kalman para ajuste de modelos lineales (Kalman, 1960), estimación de parámetros de modelos estadísticos propuesta por Akaike (Akaike, 1974), identificabilidad, estimación de parámetros y validación de modelos de espacio de estados propuesta por Ljung (Ljung, 1990; Ljung & Glad, 1994), construcción, estimación de parámetros y validación de modelos matemáticos fenomenológicos propuesta por Cameron y Hangos (Cameron & Hangos, 2001), sensibilidad de parámetros en modelos lineales de procesos biotecnológicos propuesta por Cho et al. (Cho et al., 2003), estimación de parámetros, validación y diseño óptimo de experimentos en modelos no lineales de procesos biotecnológicos propuesta por Gadkar et al. (Gadkar et al., 2005), entre otros. Sin embargo, a pesar de la variedad de métodos y metodologías propuestas para abordar los diferentes aspectos de la estimación de parámetros y la validación de modelos, se hace necesaria la construcción de una revisión sistemática de cuáles métodos aportan mejores resultados según las propiedades del sistema analizado, lo cual es de particular importancia en el caso de los procesos biotecnológicos (Barrigón et al., 2012).

En el contexto nacional, Colombia se encuentra en una posición estratégica al poseer el 10 % de la biodiversidad mundial en un área que tan solo corresponde 0.7 % de la totalidad de la superficie terrestre (Organización de las Naciones Unidas, 1992), lo que implica alta riqueza genética potencialmente utilizable. Además de lo anterior, el mercado de productos con base biotecnológica presenta un crecimiento estimado que llegara a los 727,100 millones USD en 2025 (Research G. V., 2017), junto con la expiración de aproximadamente el 50 % de las patentes farmacológicas en los próximos 10 años.

Dada la oportunidad que se presenta bajo estas condiciones, el uso de modelos matemáticos de alta calidad que permitan un mejor diseño, control y optimización en aras de obtener procesos biotecnológicos más eficientes y sostenibles en el territorio colombiano se hace imprescindible. Es por ello que a través de un caso de estudio relevante en procesos biotecnológicos como lo es la fermentación, en esta tesis se evalúan diferentes métodos en aras de consolidar una metodología sistemática para selección de métodos de tratamiento de datos, estimación de parámetros y validación de modelos matemáticos de procesos biotecnológicos. Esto con el fin de encontrar directrices que incrementen la calidad del modelo según las características del mismo, el sistema representado y los datos experimentales obtenidos.

1.1. Generalidades de modelos matemáticos

Los modelos matemáticos son relevantes debido a que permiten describir el comportamiento de los sistemas en términos de un problema matemático equivalente. Dicha representación

debe constar de un sistema de ecuaciones cuyas características básicas corresponden a una respuesta correcta en las salidas de los estados del sistema para cambios específicos en las entradas, una estructura válida que represente la conexión entre las entradas, salidas y variables internas, y una descripción correcta del comportamiento del sistema bajo diversas condiciones (Cameron & Hangos, 2001; Narayanan et al., 2020). Un tipo particular de modelos matemáticos ampliamente utilizados en procesos biotecnológicos corresponde a los modelos de caja gris de tipo fenomenológico. Estos modelos son altamente informativos al estar basados en relaciones matemáticas derivadas principios, leyes o teorías que describen los fenómenos que ocurren en el sistema físico o biológico. Las predicciones generadas con este tipo de modelos son generalmente confiables y presentan cierto grado de capacidad predictiva. Sin embargo, estos modelos incorporan parámetros cuyo valor debe ser estimado a partir de datos experimentales. La verificación de las características del modelo matemático plantea varias preguntas: ¿el modelo es identificable?, ¿qué cantidad y calidad de datos es necesaria para la estimación de parámetros?, ¿qué nivel de validación es necesario?, ¿qué nivel de precisión es adecuado?, ¿qué parámetros, entradas o perturbaciones del sistema necesitan ser conocidos para asegurar la predictibilidad del modelo?

Para contestar a estos interrogantes diversos autores han propuesto y recopilado una extensa base teórica que se divide principalmente en dos ramas: estimación de parámetros y validación de modelos, siendo el tratamiento de datos transversal a ambas (Ljung, 1990; Bastin, 2013; Englezos & Kalogerakis, 2000; Cameron & Hangos, 2001; Van den Bos, 2007). La estimación de parámetros se encarga de determinar los valores de los mismos que permiten ajustar el modelo al sistema estudiado, por otra parte, la validación de modelos verifica si verdaderamente el modelo describe el sistema cuando se cambian los parámetros de entrada del modelo (valores diferentes a los obtenidos en el ajuste).

En las secciones siguientes se explicarán brevemente las bases teóricas de la estimación de parámetros y validación de modelos junto con algunos ejemplos. Adicionalmente, se expondrán las hipótesis, objetivos, metodología y organización de la presente tesis.

1.2. Estimación de parámetros

1.2.1. Problema de optimización

La estimación de parámetros hace referencia a la determinación de valores de los parámetros de un modelo matemático, que resulta del ajuste del modelo a un conjunto de mediciones experimentales. Cuando el modelo está compuesto de ecuaciones lineales en parámetros (todos los parámetros están de la forma $a\theta$, donde θ es un parámetro y a una función de los estados del modelo) este procedimiento recibe el nombre de estimación lineal. El caso más general corresponde a la estimación no lineal, en donde las ecuaciones del modelo son funciones no

lineales en parámetros (Englezos & Kalogerakis, 2000). El caso anterior es el más común en modelos matemáticos de procesos biotecnológicos. La estimación de parámetros puede plantearse como el siguiente problema de optimización (Biegler, 2010):

$$min \ \varphi(\boldsymbol{\theta}) \tag{1-1}$$

$$s.t. \ \frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta}), \quad \mathbf{x}_0 = \mathbf{x}(t_0) \tag{1-2}$$

$$\mathbf{g}(\mathbf{x}(t), \boldsymbol{\theta}) \tag{1-3}$$

$$\mathbf{h}(\mathbf{x}(t)) \leq 0 \tag{1-4}$$

$$\boldsymbol{\theta}_L \leq \boldsymbol{\theta} \leq \boldsymbol{\theta}_U \tag{1-5}$$

$$\mathbf{x}_L \leq \mathbf{x} \leq \mathbf{x}_U \tag{1-6}$$

en donde $\varphi(\boldsymbol{\theta})$ corresponde a la función a minimizar (función objetivo), $\boldsymbol{\theta}$ vector de parámetros del modelo matemático del proceso biotecnológico, $\mathbf{f}$ vector de ecuaciones diferenciales, $\mathbf{g}$ vector de ecuaciones algebraicas, $\mathbf{h}$ vector de restricciones, $\mathbf{x}$ vector de estados, t tiempo y $\mathbf{x}_0$ vector de condiciones iniciales para las ecuaciones diferenciales. Los subindices L y U describen los valores de las cotas inferior y superior, respectivamente.

1.2.2. Función objetivo

La función objetivo $\varphi(\boldsymbol{\theta})$ es una medida de la desviación general de la salida calculada por el modelo matemático respecto a los datos experimentales. Existen diversos criterios para la selección o construcción de la función objetivo, en donde resaltan aquellos de tipo estadístico: funciones de densidad de probabilidad presentes en los datos experimentales (normal, Poisson, binomial, multinomial, etc.), máxima verosimilitud, estimadores sesgados o no sesgados, ponderación, resistencia a puntos atípicos, entre otros (Englezos & Kalogerakis, 2000; Seber & Wild, 2003; Aster et al., 2005; Prado-Rubio, 2010). Así mismo, existen otros tipos de función objetivo como las funciones regularizadas y otras con términos más complejos como la función objetivo del programa QUAL2Kw, logarítmicos, racionales, entre otros (Aster et al., 2005; Pelletier et al., 2006).

1.2.3. Solucionadores

Dependiendo de la forma en la que se proponga el modelo o se obtengan los datos experimentales, puede ser necesario un tipo particular de solucionador. Específicamente, un integrador es requerido cuando el modelo matemático es descrito en términos de un sistema de ecuaciones diferenciales ordinarias (ODE) o un sistema de ecuaciones algebraicas diferenciales (DAE) y los datos experimentales son obtenidos como valores en el tiempo (Chapra et al.,

2012). Algunos ejemplos para sistemas ODE incluyen métodos de Runge-Kutta o integradores de múltiples pasos (Chapra et al., 2012). En el caso de los sistemas DAE los integradores de tipo *Backward Differentiation Formulas* o BDF son comúnmente utilizados (Cellier & Kofman, 2006). En caso de que el modelo sea compuesto por un sistema de ecuaciones no lineales, se requiere de un solucionador de tipo "buscador de raíces". Ejemplos comunes de este tipo de algoritmos corresponden a los métodos de Newton y secante (Chapra et al., 2012).

1.2.4. Ejemplos de aplicación de estimación de parámetros

Diferentes metodologías han sido propuestas para la estimación de parámetros, aunque no siempre se ha formulado como un problema de optimización. Los parámetros de los primeros modelos de crecimiento de microorganismos para fermentaciones como la etanólica eran estimados directamente como lo reportó Sonnleitnert et al. (1986). En su investigación parámetros como la tasa de crecimiento específica máxima, la tasa de respiración específica máxima, rendimiento biomasa/glucosa, entre otros, fueron obtenidos por curvas experimentales y manipulaciones algebraicas de las ecuaciones del modelo utilizado.

Otro tipo de aproximación corresponde al uso de "observadores" o *softsensors*, los cuales son formulaciones matemáticas que permiten conocer los estados del proceso biotecnológico utilizando variables de entrada y salida del proceso medidas con un transductor (Gauthier et al., 1992). Aplicaciones prácticas de esta metodología incluyen la estimación de velocidades de reacción en biorreactores (Farza et al., 1997), estimación de las tasas de crecimiento específicas de oxidación y glucólisis de una fermentación de células animales (Dochain, 2003), parámetros del crecimiento de la bacteria *Bacillus thurigiensis* (Bogaerts & Wouwer, 2003), entre otros.

Los observadores también han servido para estimación de parámetros de crecimiento de *Escherichia coli* en la producción de ampicilina (Roman & Selişteanu, 2012), estimación de la tasa especifica de consumo de sustrato y control simultáneos de una fermentación de ácido láctico (Sagmeister et al., 2013) y estimación de parámetros cinéticos en producción de lipasas con la bacteria *Candida rugosa* (Selişteanu et al., 2014). Un nuevo enfoque dado en la investigación del uso de observadores corresponde a la comparación de múltiples formulaciones de los mismos, que se encargan de mejorar aspectos como la no linealidad presente en los modelos de procesos biotecnológicos (Bastin & Dochain, 1990). Ejemplos de este enfoque incluyen la estimación simultánea de condiciones iniciales y parámetros, aplicada al cultivo de células humanas embrionales de riñón para la estimación de tasas específicas de crecimiento y concentración de biomasa inicial (Hulhoven et al., 2008) y la estimación de tasas de crecimiento específico en producción de proteínas heterólogas con la bacteria *Pichia pastoris* (Barrigón et al., 2012).

Las metodologías anteriores estiman los parámetros desconocidos de los modelos matemáticos de una manera directa a medida que se resuelven las ecuaciones diferenciales, por lo tanto, no hacen uso de optimización. Lo anterior es verificable al observar que los parámetros obtenidos con observadores son variables en el tiempo, en contraste con el problema de optimización en donde los mismos se definen como invariantes en el tiempo.

Cuando los parámetros son estimados a través de un proceso de optimización, la función objetivo se deriva de un sistema de ecuaciones altamente no lineal, lo que conduce al problema de múltiples óptimos locales o problema no convexo (Biegler, 2010). Para encontrar el mínimo global de la función objetivo surgen dos enfoques: optimización determinista y optimización estocástica. La primera hace referencia a algoritmos que cuentan con una programación matemática rígida normalmente basada en el gradiente de la función objetivo y no cuenta con elementos aleatorios, sin embargo, puede presentar problemas en mínimos locales. La segunda corresponde al uso de algoritmos que presentan elementos aleatorios y es comúnmente utilizada en optimización global (Cavazzuti, 2012).

El campo de la optimización determinística algunos algoritmos como Newton, Cuasi-Newton, interior point, simplex y Levenberg-Marquardt son ampliamente utilizados, siendo el último aplicado en el modelamiento de recuperación de una enzima hidrogenasa modificada producida en *Saccharomyces cerevisiae* (Varga et al., 2001), estimación de parámetros cinéticos y control de la producción del aminoácido metionina en fed-batch en conjunto con algoritmos de redes neuronales (Nayak & Gomes, 2009), estimación de parámetros cinéticos en crecimiento diaúxico de galactosa y acetato en la bacteria *Salmonella typhimurium*, aproximación de cinéticas de reacción por *splines* en bioprocesos (Mašić et al., 2017), entre otros.

Por otro lado, la optimización estocástica se basa principalmente en algoritmos que representan reglas heurísticas tomadas de la naturaleza, entre ellas templado simulado (simulated anneling), enjambre de partículas, teoría de juegos y algoritmos evolutivos (Cavazzuti, 2012). Este último se basa en la teoría evolutiva de Darwin (Darwin, 2004) tomando el vector de parámetros como un individuo y seleccionando individuos según el desempeño de la función objetivo (Koza, 1990). Este algoritmo fue utilizado para estimación de parámetros en el proceso fermentativo de avermectina, una sustancia insecticida, acaricida y antihelmíntica (Wu et al., 2005). Se han reportado metodologías híbridas que combinan algoritmos deterministas y estocásticos, en donde el determinista calcula un estimado inicial y el estocástico está dado por *cadenas de Markov*, en este caso para la estimación de parámetros en cinéticas enzimáticas generalizadas (Grosfils et al., 2007).

1.3. Calidad de predicción

1.3.1. Validación de modelos matemáticos

La validación de un modelo matemático corresponde al proceso de determinar si las variables de salida del modelo que representa un sistema real (en este caso un proceso biotecnológico), se encuentra dentro de alguna tolerancia (National Research Council, 2012). La validación de un modelo matemático puede ser evaluada con respecto a conocimiento previo, comportamiento del modelo en simulaciones y datos experimentales (Ljung, 1990; Consonni et al., 2010; Henninger et al., 2010; Keesman, 2011; Ling & Mahadevan, 2013; Sargent, 2013; Rajamanickam et al., 2021). Un concepto relevante en la validación de modelos matemáticos corresponde al valor "verdadero" del parámetro. Este concepto proviene desde la estadística y hace referencia al valor que tendría un parámetro si el mismo fuera calculado utilizando todos los elementos de una población, por lo que carecería de incertidumbre (Kosorok, 2008). Generalmente, solo es posible utilizar una muestra de los elementos pertenecientes a la población, por lo que los parámetros estimados tienen algún grado de incertidumbre. De forma particular para modelos matemáticos fenomenológicos, un concepto relevante es el de interpretabilidad de parámetros. Un parámetro es interpretable cuando tiene significado físico y magnitud (o valor) conocida (Lema-Perez et al., 2019). Sin embargo, también se requiere que la magnitud del parámetro pueda ser estimada con una baja incertidumbre.

- **Validación por conocimiento previo:**

  Una primera prueba realizada para conocer si un modelo describe apropiadamente un sistema corresponde a la evaluación de los valores de los parámetros obtenidos en la estimación. En sistemas físicos como los representados por procesos biotecnológicos, se espera que los parámetros sean positivos. El caso contrario significa que hubo algún problema en la estimación de parámetros, normalmente, que no se alcanzó el mínimo global de la función objetivo (Keesman, 2011). Este tipo de validación es posible gracias a la interpretabilidad del parámetro y su magnitud, que permite establecer un rango "lógico" acorde a límites físicos del fenómeno representado.

  Un aspecto importante de los parámetros estimados que es útil en validación de modelos corresponde a la varianza de estimación de dichos parámetros (σ_θ^2). La varianza se define como el cuadrado de la desviación del parámetro respecto a su valor medio, tomando en cuenta que el valor estimado es el valor medio del parámetro (Van den Bos, 2007). La matriz de información de Fisher F expresa la matriz de covarianza de los parámetros estimados y se define como:

$$F = E\left[s_\theta s_\theta^T\right] = E\left[\frac{\partial q(\mathbf{x};\boldsymbol{\theta})}{\partial \theta}\frac{\partial q(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] \tag{1-7}$$

en donde $\boldsymbol{\theta}$ corresponde al vector de parámetros y $q(\mathbf{x};\boldsymbol{\theta})$ a los valores predichos por el modelo dado un vector de parámetros estimados. El grado de certeza con que se conoce el valor estimado de un parámetro está relacionado con su desviación estándar (σ_θ), siendo mayor a medida que la desviación estándar tiende a cero. Lo anterior implica el valor "verdadero" de un parámetro se encuentra acotado en un rango menor, y por ende, el valor estimado es más confiable (Van den Bos, 2007). La desviación estándar de los parámetros estimados puede conocerse a partir de F:

$$\sigma_{\boldsymbol{\theta}} = \sqrt{diag(F)} = (\sigma_{\boldsymbol{\theta}1}, \ldots, \sigma_{\boldsymbol{\theta}N}) \tag{1-8}$$

- **Validación por experiencia con el modelo:**

Este tipo de validación busca evaluar la respuesta del modelo con los parámetros estimados mediante simulación. Dado un entendimiento de las trayectorias que sigue el sistema representado por el modelo matemático bajo condiciones normales, se examina si existen discrepancias entre la salida del modelo y la trayectoria esperada (Keesman, 2011). Lo anterior puede observarse en la Figura **1-1**, que muestra las trayectorias para un modelo matemático de consumo de sustrato por parte de un microorganismo, descrito a continuación:

$$S(t) = S(t-1) - \mu \frac{S(t-1)}{K_s + S(t-1)},\ S(0) = 30\,\mathrm{g/m^3} \tag{1-9}$$

en donde t es el tiempo, S concentración de sustrato en $\mathrm{g/m^3}$, μ constante de crecimiento específica máxima en $\mathrm{min^{-1}}$, K_s constante de saturación en $\mathrm{g/m^3}$. La línea roja muestra la trayectoria obtenida con parámetros "verdaderos", en tanto que la línea azul muestra aquella obtenida con parámetros estimados. Claramente se observa una discrepancia debida al valor negativo en el parámetro K_s.

- **Calidad de predicción con datos experimentales:**

Sumada a la validación por conocimiento previo y experiencia con el modelo, puede realizarse otro tipo de validación con el uso de los errores e, también llamados residuales. Estos poseen una importancia clave en el proceso de validación al representar la desviación de la predicción del modelo con la medición experimental (Keesman, 2011). Existen algunos tipos comunes de pruebas de residuales como:

  - ◇ **Inspección gráfica:** consiste en representar gráficamente los residuales como una función del tiempo. Esta representación puede mostrar directamente tendencias como deriva, periodicidad y puntos atípicos (Keesman, 2011).

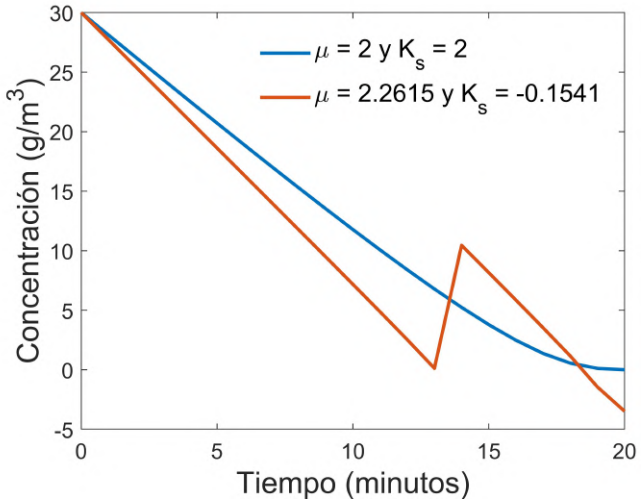


Figura 1-1: Validación por experiencia con el modelo

◇ **Pruebas de residuales:**

∗ **Prueba de Blancura de residuales o autocorrelación:** idealmente los residuales del modelo matemático no deben depender de residuales previos, es decir, deben ser independientes (Keesman, 2011). La función de autocorrelación está dada por:

$$r_{ee}(l) = \frac{1}{N-l}\sum_{i=1}^{N-1} e_i e_{i+l} \tag{1-10}$$

en donde N es el número de datos experimentales, e residual del modelo, i y l contadores. Para probar si los residuales son independientes entre sí y están normalmente distribuidos puede usarse la siguiente expresión:

$$\frac{N}{\hat{r}_{ee}(0)^2}\sum_{l=1}^{M}\hat{r}_{ee}(l)^2 \leq X_\alpha^2(M) \tag{1-11}$$

en donde $X_\alpha^2(M)$ corresponde al valor de la distribución chi-cuadrado para M grados de libertad y un nivel de significatividad α. Si dicha igualdad se mantiene, los residuales son independientes entre sí para un rango M de datos experimentales.

∗ **Prueba de correlación cruzada:** en este caso, se comprueba la correlación de los residuales e con las variables de entrada u, según la función:

$$r_{ue}(l) = \frac{1}{N-l}\sum_{i=1}^{N-1} u_i e_{i+l} \tag{1-12}$$

Para determinar estadísticamente que los residuales y las entradas son independientes entre sí o no correlacionadas, se utiliza la fórmula:

$$|\hat{r}_{ue}(l)| \leq \sqrt{\frac{\sum_{i=-\infty}^{\infty} r_{ee}(l) r_{uu}(l)}{N}} Norm_{\alpha} \tag{1-13}$$

en donde r_{ee} y r_{uu} corresponden a las funciones de autocorrelación de residuales y entradas, respectivamente. $Norm$ es el valor de la distribución normal estándar con un nivel de significancia α.

◇ **Intervalos de predicción:** es posible calcular los intervalos de predicción de la salida del modelo para un nivel de precisión especificado, es decir, conocer el intervalo de valores en donde puede existir una predicción del modelo realizada con los parámetros "verdaderos" (Cappuyns et al., 2009). El intervalo de predicción se calcula como:

$$\mathbf{x}_i \pm t_{(1-\alpha/2, N-p)} \sqrt{S^2_{\mathbf{x_i}(\boldsymbol{\theta})} + S^2_{y_i}} \tag{1-14}$$

$$\left[\frac{\partial \mathbf{f}_i}{\partial \boldsymbol{\theta}}\right] \cdot \boldsymbol{C_{\theta}} \cdot \left[\frac{\partial \mathbf{f}_i}{\partial \boldsymbol{\theta}}\right]^T \tag{1-15}$$

en donde $\mathbf{x}_i$ es la predicción del modelo para el estado i, $t_{(1-\alpha/2, N-P)}$ es el valor del estadístico t de Student para una confianza de $1-\alpha/2$ y grados de libertad $N-p$, N número de mediciones experimentales, p número de parámetros, $S^2_{\mathbf{x_i}(\boldsymbol{\theta})}$ varianza del error de predicción, $S^2_{y_i}$ varianza de medición experimental, $\partial \mathbf{f}_i / \partial \boldsymbol{\theta}$ sensibilidad de salidas del modelo respecto a parámetros y $\boldsymbol{C_{\theta}}$ matriz de covarianza de los parámetros.

∗ **Indicadores de desempeño:** calcular cuantitativamente la desviación de la salida del modelo respecto a los datos experimentales es una aproximación útil en la comparación de diferentes conjuntos de parámetros obtenidos por estimación de parámetros (Barrigón et al., 2012). Algunos de estos indicadores son:

∗ Suma de cuadrados del error (SSE):

$$SSE = \sum_{i=1}^{N} (\mathbf{x}_i - y_i)^2 \tag{1-16}$$

* Error cuadrático medio (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\mathbf{x}_i - y_i)^2}{N}} \tag{1-17}$$

* Error relativo medio (MRE):

$$MRE = \frac{1}{N}\sum_{i=1}^{N}\frac{|(\mathbf{x}_i - y_i)|}{y_i} \tag{1-18}$$

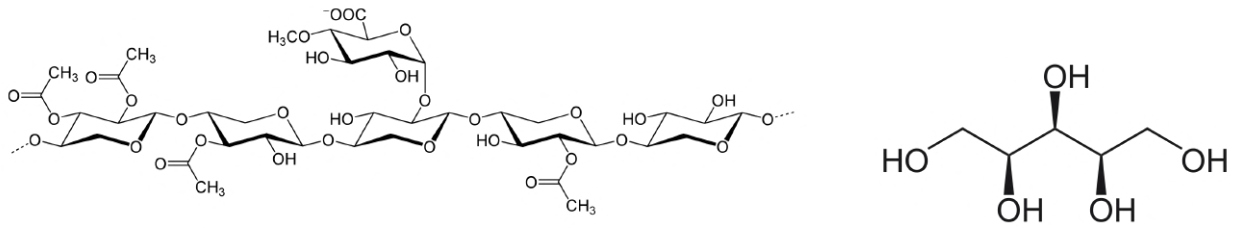
1.3.2. Ejemplos de aplicación de validación de modelos

Desde los inicios de la simulación de procesos biotecnológicos, el tipo más común de validación que se ha realizado corresponde a una "curva de validación" o inspección gráfica. En dicho caso se representan simultáneamente los puntos experimentales y la predicción del modelo, en donde se espera que la predicción siga la tendencia general de los puntos experimentales y a su vez, se aproxime a estos. La curva de validación ha sido aplicada en fermentaciones de penicilina en conjunto con redes neuronales (Di Massimo et al., 1992) y ecuaciones diferenciales (van Can et al., 1999), optimización de producción de vino con redes neuronales (Vlassides et al., 2001), producción de la enzima β-glucuronidasa en bacterias transgénicas (Sun et al., 2001), análisis de inactivación térmica de enzimas (Illanes & Wilson, 2003), biorremediación de hexano en suelo (Schoefs et al., 2003), análisis de genes represores en fermentación etanólica (Kobayashi & Nakamura, 2004), secreción de proteínas heterólogas en levaduras (Rakestraw et al., 2006), producción de enzimas recombinantes en bacterias (Nadri et al., 2006), distribución de factores de transcripción en células sanguíneas (Csaszar et al., 2009), producción de células de embrión (Yeo et al., 2013), reactores de membrana biocatalíticos (du Preez et al., 2015), estimación de parámetros de un modelo de fermentación ABE (Acetona-Butanol-Etanol) (Díaz & Willis, 2018), entre otros.

Otras metodologías usadas en la validación de modelos de procesos biotecnológicos se basan en indicadores de desempeño. Uno de estos indicadores comúnmente utilizado corresponde al cálculo del error de ajuste. Dentro de esta categoría se encuentra el indicador de suma de cuadraros del error (Stigler, 1981) que a su vez ha sufrido diversas modificaciones. Ejemplos de este criterio incluyen fermentación etanólica (Wang et al., 2001) y producción de goma xantana (Zabot et al., 2011). Otro criterio de error de ajuste utilizado es el error relativo absoluto promedio, aplicado a estimación de biomasa por fluorescencia (Surribas et al., 2006), fermentación láctica (Zhang et al., 2011) y degradación anaerobia de lodos de aguas residuales (Donoso-Bravo et al., 2014). El error relativo medio fue utilizado como criterio en un modelo poblacional de células suspendidas (Kolewe et al., 2011).

1.4. Caso de estudio: bioproducción de xilitol

El xilitol es un alcohol derivado de la xilosa, azúcar de 5 carbonos o pentosa, que se encuentra en forma del polímero conocido como hemicelulosa o xilano presente en el tejido vegetal (Figura **1-2**). El xilitol es una molécula altamente versátil con una variedad de usos que abarcan las cremas dentales, edulcorantes, medicamentos, suplementos dietarios e incluso ha sido seleccionado como una plataforma química para la producción de otros compuestos ya sea por vía química o biotecnológica (cofermentación) (Werpy & Petersen, 2004). En cuanto al mercado mundial de xilitol este alcanzó un volumen de 190.9 millones de toneladas valoradas en 726 millones de dólares en 2016, en donde se estima una proyección a un volumen de producción de 266.6 millones de toneladas valoradas en 1000 millones de dólares a 2022 (Ahuja & Mamtani, 2019).



(a) Xilano (b) Xilitol

Figura 1-2: Estructura química del (a) xilano y (b) xilitol (Arcaño et al., 2020).

Diferentes estudios se han realizado sobre la bioproducción de xilitol a partir de recursos renovables, entre estas materias primas se incluye el bagazo de caña de azúcar (concentración final de 36.11 g/L), cáscaras de banano (concentración final de 24.7 g/L) y raquis de maíz (concentración final de 35 g/L) (Hernández-Pérez et al., 2016; Rehman et al., 2018; Kumar et al., 2019). Así mismo, diferentes tipos y especies de microorganismos han sido utilizados para la producción de este metabolito como bacterias (*Enterobacter liquefaciens*, *Corynebacterium sp.*), hongos filamentosos (*Penicillium crustosom*, *Aspergillus niger*) y levaduras (*Candida guilliermondii*, *Candida tropicalis*, *Candida sp.*, *Pichia sp.*, *Pichia stipitis*, *Debaromyces hansenii*, *Debaromyces nepalensis*, *Hansunela anomala*, *Kluyveromyces marxianus*), entre otras (Dasgupta et al., 2017). Desde el punto de vista del proceso industrial, la separación y purificación del xilitol obtenido por bioproducción son un reto debido la mezcla compleja resultante del medio de fermentación. Diferentes alternativas han sido investigadas como carbón activado (80 % de recuperación), resinas de intercambio iónico (34 % de recuperación), extracción líquido-líquido (21.72 g/L extraídos), membranas (85 % de recuperación), precipitación (64 % de recuperación), métodos combinados (34 % de recuperación), entre otros (Martínez et al., 2015). Desde el punto de vista de modelamiento matemático, se ha estudiado la optimización de la producción de xilitol en fermentaciones de tipo fed-batch (Prado-Rubio et al., 2015; Hernández-Escoto et al., 2016).

Como modelo matemático de prueba para demostrar la aplicación de la metodología investigada se usa el modelo de fermentación diaúxica de glucosa y xilosa para producción de xilitol propuesto por Tochampa et al. (2015). Este modelo se compone por seis ecuaciones diferenciales correspondientes a los estados del sistema, 16 parámetros (5 especificados, 11 desconocidos) y seis ecuaciones adicionales. Además, el modelo matemático presenta como características (i) inhibición en la recepción de sustrato por efecto antagónico de glucosa y xilosa, (ii) transporte de xilitol desde la célula hacia el exterior y (iii) consideración de la estequiometría molar de la reacción de producción de xilitol a partir de xilosa. Las variables y parámetros del modelo están descritas en la Tabla **1-1**. Las ecuaciones y relaciones adicionales son mostradas a continuación:

Modelo dinámico:

$$\frac{dC_X}{dt} = -\frac{F_{glu}+F_{xil}}{V_L}C_X + \mu C_x \tag{1-19}$$

$$\frac{dC_{xil}}{dt} = \frac{F_{xil}}{V_L}C^f_{xil} - \frac{F_{glu}+F_{xil}}{V_L}C_{xil} - q_{xil}C_X \tag{1-20}$$

$$\frac{dC_{glu}}{dt} = \frac{F_{glu}}{V_L}C^f_{glu} - \frac{F_{glu}+F_{xil}}{V_L}C_{glu} - q_{glu}C_X \tag{1-21}$$

$$\frac{dC^{ex}_{xit}}{dt} = -\frac{F_{glu}+F_{xil}}{V_L}C^{ex}_{xit} - r^{'}_{t,xit}C_X \tag{1-22}$$

$$\frac{dC^{in}_{xit}}{dt} = (r_{f,xit} - r_{u,xit} - r^{'}_{t,xit})\rho_X - \mu C^{in}_{xit} \tag{1-23}$$

$$\frac{dV_L}{dt} = F_{xil} + F_{glu} \tag{1-24}$$

Expresiones cinéticas:

$$\mu = \mu^{max}_{glu}\frac{C_{glu}}{K_{S,glu}+C_{glu}} + \mu^{max}_{xit}\frac{C^{in}_{xit}}{K_{S,xit}+C^{in}_{xit}}\frac{K_r}{K_r+C_{glu}} \tag{1-25}$$

$$q_{glu} = q^{max}_{glu}\frac{C_{glu}}{C_{glu}+K_{S,glu}\left(1+\frac{C_{xil}}{K_{i,xil}}\right)} \tag{1-26}$$

$$q_{xil} = q^{max}_{xil}\frac{C_{xil}}{C_{xil}+K_{S,xil}\left(1+\frac{C_{glu}}{K_{i,glu}}\right)} \tag{1-27}$$

$$r^{'}_{t,xil} = 3.6\text{x}10^6 P_{xit}a_{cel}(C^{in}_{xit} - C^{ex}_{xit}) \tag{1-28}$$

$$r_{f,xit} = \frac{M_{xit}}{M_{xil}}q_{xil} \tag{1-29}$$

$$r_{u,xit} = \frac{\mu_{xit}}{Y_{X/xit}} \tag{1-30}$$

Tabla 1-1: Parámetros y variables del modelo de producción diaúxica de xilitol (Tochampa et al., 2015).

Variables		
Símbolo	Descripción	Unidad
C_X	Concentración de biomasa	g/L
C_{xil}	Concentración de xilosa	g/L
C_{glu}	Concentración de glucosa	g/L
C_{xit}^{ex}	Concentración de xilitol extracelular	g/L
C_{xit}^{in}	Concentración de xilitol intracelular	g/L
V_L	Volumen de fermento	L
Parámetros		
Símbolo	Descripción	Unidad
μ_{glu}^{max}	Tasa máxima de crecimiento en glucosa	h^{-1}
μ_{xit}^{max}	Tasa máxima de crecimiento en xilitol	h^{-1}
$K_{S,glu}$	Constante de saturación en glucosa	g/L
$K_{S,xit}$	Constante de saturación en xilitol	g/L
$K_{S,xil}$	Constante de saturación en xilosa	g/L
K_r	Constante de represión por glucosa	g/L
q_{xil}^{max}	Tasa de ingreso máximo de xilosa	g xilosa/DCW h
q_{glu}^{max}	Tasa de ingreso máximo de glucosa	g glucosa/DCW h
$K_{i,xil}$	Constante de inhibición por xilosa	g/L
$K_{i,glu}$	Constante de inhibición por glucosa	g/L
P_{xit}	Coeficiente de permeabilidad de la membrana celular a xilitol	m/s
ρ_X	Densidad celular	120 g DCW/L
a_{cel}	Área superficial específica de la célula	7.6 m^2/g
M_{xit}	Peso molecular del xilitol	152 g/mol
M_{xil}	Peso molecular de la xilosa	150 g/mol
$Y_{X/xit}$	Rendimiento de biomasa en xilitol	0.48 g biomasa/g xilitol

1.5. Motivación, hipótesis y objetivos

1.5.1. Motivación

Los modelos matemáticos han representado una parte fundamental en la ingeniería de procesos, pues los mismos son necesarios para su diseño, control y optimización. Actualmente, existe además un movimiento de transición hacia nuevas tecnologías caracterizadas por una mayor responsabilidad tanto social como ambiental. Entre ellas, se encuentra la biotecnología como alternativa que brinda oportunidades para reemplazar las tecnologías de producción

basadas en combustibles fósiles, por tecnologías basadas en sistemas biológicos o partes de los mismos. Sin embargo, estos sistemas son altamente complejos y poseen una variabilidad intrínseca que dificulta su manipulación. Este punto es donde los modelos matemáticos de procesos biotecnológicos hacen un valioso aporte, permitiendo un mayor entendimiento de la fenomenología subyacente en los sistemas biológicos.

No obstante, la variabilidad inherente a estos sistemas y sus particularidades deben no solo ser capturadas por el modelo matemático, sino también, descritas y predichas por el mismo. Los parámetros son entonces el medio por el cual el modelo describe a un sistema biológico específico. Llegar a conocer el valor de estos parámetros se convierte en una tarea compleja que involucra la mezcla de conocimiento teórico (fenomenología) y comportamiento real (datos experimentales) del sistema, junto con herramientas matemáticas (modelamiento matemático, optimización, estadística). Adicionalmente, los elementos ya mencionados involucrados en la estimación de parámetros de modelos matemáticos de procesos biotecnológicos junto con sus interacciones pueden afectar la calidad del modelo matemático. Por esta razón, se hace necesaria robustecer las aproximaciones sistemáticas (en métodos y metodologías) que permitan incrementar la interpretabilidad en los parámetros y mejorar la calidad de la predicción del modelo. Esta tesis de investigación busca entonces consolidar una metodología que aborde de manera sistemática la estimación de parámetros en modelos matemáticos de procesos biotecnológicos, cuya aplicación permita obtener modelos con una alta capacidad descriptiva y predictiva.

1.5.2. Hipótesis

Una mejor calidad de los modelos matemáticos utilizados en el diseño, control y optimización de procesos biotecnológicos puede llevar a un mayor entendimiento de los mismos, y por tanto, oportunidades para uso u optimización. Debido a lo anterior, las hipótesis bajo las cuales se realizará esta investigación son:

- Es posible la generación de directrices para estimación de parámetros con base en las características del modelo y sistema, que mejoren la significancia de los parámetros.
- Es posible la generación de directrices para selección de métodos de validación con base en las características del modelo y sistema, que permitan evaluar cuantitativamente la calidad del modelo.

1.5.3. Objetivos

Objetivo general:

Generar directrices para la selección de métodos de estimación de parámetros y validación de modelos que permitan un mayor grado de significancia de parámetros y una mayor calidad

del modelo.

Objetivos específicos:

- Seleccionar casos de estudio relevantes para estimación de parámetros en modelos de procesos biotecnológicos.
- Investigar métodos de estimación de parámetros y cuantificar su eficiencia.
- Investigar métodos de validación de modelos con base en la calidad predictiva del modelo.
- Formular una metodología de estimación de parámetros y validación de modelos.

1.6. Organización de la tesis

La metodología presentada por la Figura **1-3** es un esquema general de los pasos utilizados en esta investigación, que sirve como guía al lector y muestra cómo se interconectan los distinto elementos analizados en esta tesis. Los detalles específicos de cada proceso se especifican en los capítulos correspondientes, dado que cada capítulo representa un paso en la construcción de la metodología final.

Esta tesis está organizada en capítulos escritos en forma de artículo científico, y por lo tanto, se anticipa al lector que parte de la información se encontrará reiterativamente a lo largo del documento (principalmente el modelo matemático de caso de estudio). Esto se hace con el objetivo de brindarle coherencia a las contribuciones de manera independiente.

El capítulo 1 titulado "Introducción" expone nociones básicas de estimación de parámetros y validación de modelos, junto con las hipótesis, objetivos, metodología y organización de la presente tesis.

El capítulo 2 titulado "Tratamiento de datos para estimación de parámetros" describe las manifestaciones de error en las mediciones experimentales, los efectos que esto puede ocasionar en el proceso de estimación de parámetros y métodos para eliminar dichos errores. Metodológicamente se analizan algoritmos de detección de puntos atípicos, reemplazo de puntos atípicos y suavizado de datos para eliminación de error aleatorio en datos experimentales reales.

El capítulo 3 titulado "Identificabilidad estructural de parámetros" presenta la identificabilidad estructural de parámetros, es decir, si el valor de un parámetro del modelo matemático puede ser conocido a partir de ciertas mediciones experimentales. Se hace una descripción de las aproximaciones más utilizadas y el software disponible, así como los resultados obtenidos

para el modelo de caso de estudio. La contribución "Análisis preliminares e identificabilidad práctica en modelos de procesos biotecnológicos" hace parte de este capítulo y fue escrita como parte del XL ENCUENTRO NACIONAL DE LA AMIDIQ, además, contó con una presentación de tipo poster.

El capítulo 4 titulado "Sintonización de algoritmos de optimización" presenta sintonización de los algoritmos de optimización incluidos en el *Global optimization toolbox* de Matlab® R2018 a través de interconexión con el software R® v4.0.3. Así mismo, se expone el análisis de los efectos del tipo de normalización de función objetivo, algoritmo de optimización y sintonización del optimizador en la reproducibilidad y precisión del valor calculado para la función objetivo del presente caso de estudio. Esta contribución se presentó como artículo científico cuyo manuscrito se encuentra en proceso de ser sometido a la revista Biotechnology Journal.

El capítulo 5 titulado "Estimación práctica de parámetros" reúne los diferentes elementos que conforman el problema de optimización analizados en capítulos anteriores (modelo matemático, función objetivo, solucionadores, datos experimentales y algoritmo de optimización) y presenta su solución junto la determinación de las capacidades descriptiva y predictiva del caso de estudio a través de pruebas estadísticas, indicadores de ajuste e índices de incertidumbre. Este capítulo presenta entonces la consolidación de la metodología propuesta en esta investigación. Esta contribución corresponde a un artículo científico en construcción. Adicionalmente, la contribución "Assessing parameter relative importance in bioprocesses mathematical models through dynamic sensitivity analysis" hace parte de este capítulo, resultado de una estancia internacional en la Universidad de Guanajuato. Este artículo fue escrito para la 30va versión del European Symposium on Computer Aided Process Engineering - ESCAPE, además, contó con una presentación oral.

Finalmente, el capítulo 6 titulado "Conclusiones y perspectivas" presenta una visión general de las contribuciones al conocimiento realizadas en esta tesis junto con sugerencias para trabajos futuros que complementen esta investigación.

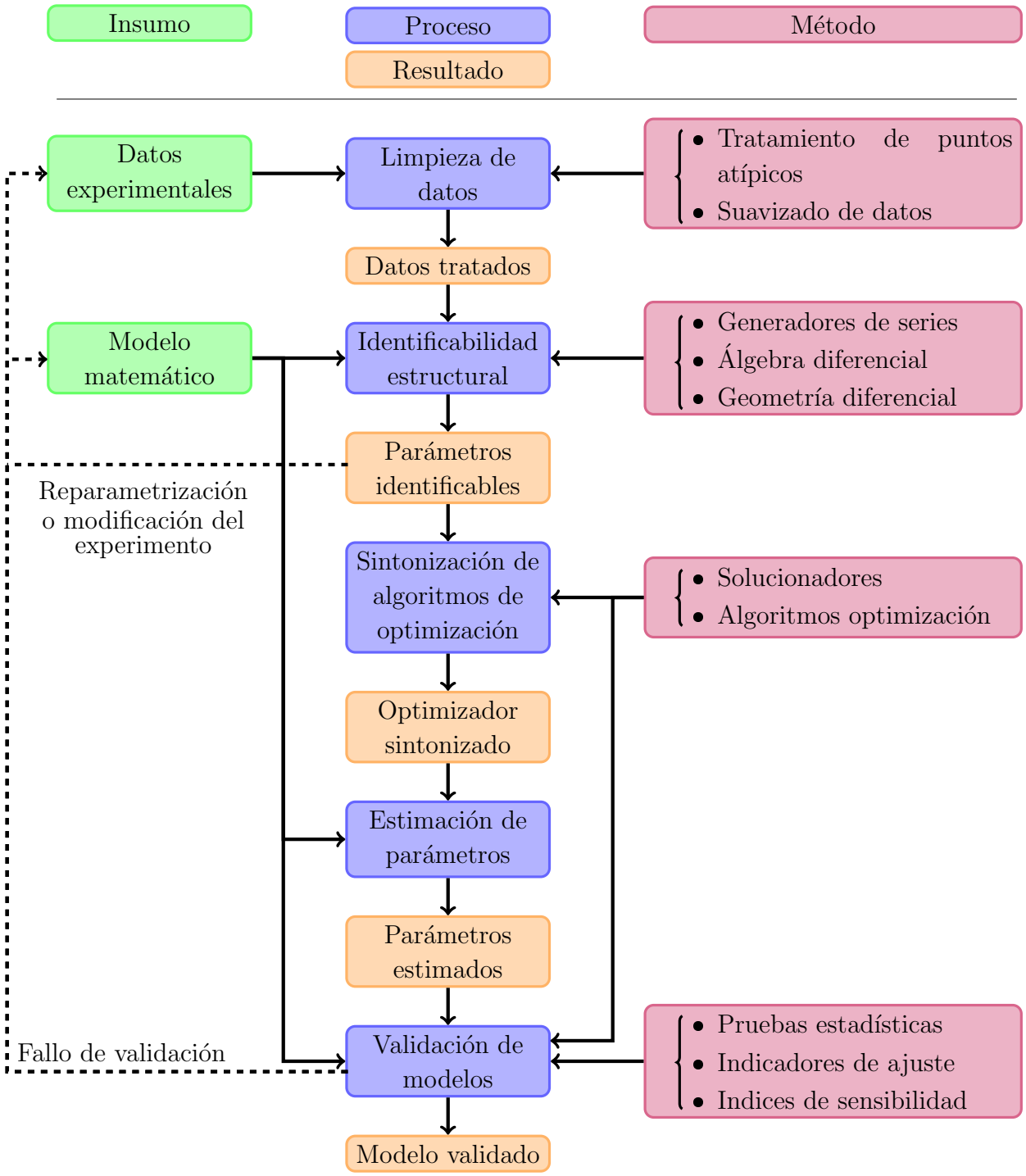


Figura 1-3: Metodología general.

1.7. Contribuciones

Las contribuciones aportadas por esta tesis de investigación son:

- Artículos científicos sometidos:
  - "Assessment of Metaheuristic-Optimization Algorithms Tuning for Parameter Estimation of Xylitol Fermentation Kinetics" sometido a la revista Biotechnology Journal, ISSN: 1860-7314.
- Artículos científicos en construcción:
  - "Parameter estimation and model validation in bioprocess mathematical models: an extensive methodology" se encuentra en construcción con avance del 90 %.
- Artículos peer reviewed:
  - Artículo de conferencia: Sánchez-Rendón, J. C., Morales-Rodriguez, R., Matallana-Pérez, L. G., & Prado-Rubio, O. A. (2020). Assessing Parameter Relative Importance in Bioprocesses Mathematical Models through Dynamic Sensitivity Analysis. In Computer Aided Chemical Engineering (Vol. 48, pp. 1711-1716). doi: 10.1016/B978-0-12-823377-1.50286-X
  - Artículo de conferencia: "Análisis preliminares e identificabilidad práctica en modelos de procesos biotecnológicos" publicado en *MEMORIAS XL ENCUENTRO NACIONAL DE LA AMIDIQ*, páginas: 3641-3646, ISBN: en trámite
- Presentaciones en eventos:
  - Presentación en modalidad oral: "Assessing parameter relative importance in bioprocesses mathematical models through dynamic sensitivity analysis", en el *30th European Symposium on Computer Aided Process Engineering (2020)*, Milán, Italia
  - Modalidad poster: "Análisis preliminares e identificabilidad práctica en modelos de procesos biotecnológicos", en el *XL ENCUENTRO NACIONAL DE LA AMIDIQ (2019)*, Huatulco, México
- Estancia internacional:
  - A cargo del profesor Ph.D. Ricardo Morales Rodríguez, Coordinador de posgrado, Programa de Ingeniería Química. Universidad de Guanajuato.
  - Duración: 2 meses.

Bibliografía

Ahuja, K. & Mamtani, K. (2019). Xylitol Market Size By Application (Chewing Gum, Confectionary, Food, Personal Care, Pharmaceuticals, Nutraceuticals), Downstream Application Potential (Xylaric Acid, Ethylene Glycol, Propylene Glycol), Industry Analysis Report, Regional Outlook, Industry Outlook Report, Regional Analysis, Application Potential, Price Trends, Competitive Market Share & Forecast, 2020 - 2026. Technical report, Global Market Insights.

Akaike, H. (1974). A new look at the statistical model identification. IEEE transactions on automatic control, 19(6), 716–723.

Arcaño, Y. D., García, O. D. V., Mandelli, D., Carvalho, W. A., & Pontes, L. A. M. (2020). Xylitol: A review on the progress and challenges of its production by chemical route. Catalysis Today, 344, 2–14.

Aster, R. C., Borchers, B., & Thurber, C. H. (2005). Parameter estimation and inverse problems. Elsevier.

Banga, J. R., Moles, C. G., & Alonso, A. A. (2004). Global optimization of bioprocesses using stochastic and hybrid methods. In Frontiers in global optimization (pp. 45–70). Springer.

Barrigón, J. M., Ramon, R., Rocha, I., Valero, F., Ferreira, E. C., & Montesinos, J. L. (2012). State and specific growth estimation in heterologous protein production by pichia pastoris. AIChE journal, 58(10), 2966–2979.

Bastin, G. (2013). On-line estimation and adaptive control of bioreactors, volume 1. Elsevier.

Bastin, G. & Dochain, D. (1990). On-line estimation and adaptive control of bioreactors, volume 1. Elsevier.

Biegler, L. T. (2010). Nonlinear programming: concepts, algorithms, and applications to chemical processes, volume 10. Siam.

Bogaerts, P. & Wouwer, A. V. (2003). Software sensors for bioprocesses. ISA transactions, 42(4), 547–558.

Brun, R., Reichert, P., & Künsch, H. R. (2001). Practical identifiability analysis of large environmental simulation models. Water Resources Research, 37(4), 1015–1030.

Cameron, I. T. & Hangos, K., Eds. (2001). Process Modelling and Model Analysis, volume 4. Academic press.

Cappuyns, A. M., Bernaerts, K., Vanderleyden, J., & Van Impe, J. F. (2009). A dynamic model for diauxic growth, overflow metabolism, and ai-2-mediated cell–cell communication of salmonella typhimurium based on systems biology concepts. Biotechnology and bioengineering, 102(1), 280–293.

Cavazzuti, M. (2012). Optimization methods: from theory to design scientific and technological aspects in mechanics. Springer Science & Business Media.

Cellier, F. E. & Kofman, E. (2006). Differential algebraic equation solvers. In Continuous system simulation (pp. 319 – 396).: Springer.

Chapra, S. C. et al. (2012). Applied numerical methods with MATLAB for engineers and scientists. New York: McGraw-Hill,.

Cho, K.-H., Shin, S.-Y., Kolch, W., & Wolkenhauer, O. (2003). Experimental design in systems biology, based on parameter sensitivity analysis using a monte carlo method: A case study for the tnfα-mediated nf-κ b signal transduction pathway. Simulation, 79(12), 726–739.

Comisión Europea (2010). Europa 2020: Una estrategia para un crecimiento inteligente, sostenible e integrador. Technical report.

Comisión Europea (2018). Un planeta limpio para todos: la visión estratégica europea a largo plazo para una economía próspera, moderna, competitiva y climáticamente neutra. Technical report.

Consonni, V., Ballabio, D., & Todeschini, R. (2010). Evaluation of model predictive ability by external validation techniques. Journal of chemometrics, 24(3-4), 194–201.

Csaszar, E., Gavigan, G., Ungrin, M., Thérien, C., Dubé, P., Féthière, J., Sauvageau, G., Roy, D. C., & Zandstra, P. W. (2009). An automated system for delivery of an unstable transcription factor to hematopoietic stem cell cultures. Biotechnology and Bioengineering, 103(2), 402–412.

Darwin, C. (2004). On the origin of species, 1859. Routledge.

Dasgupta, D., Bandhu, S., Adhikari, D. K., & Ghosh, D. (2017). Challenges and prospects of xylitol production with whole cell bio-catalysis: A review. Microbiological research, 197, 9–21.

Di Massimo, C., Montague, G., Willis, M., Tham, M., & Morris, A. (1992). Towards improved penicillin fermentation via artificial neural networks. Computers & Chemical Engineering, 16(4), 283–291.

Díaz, V. H. G. & Willis, M. J. (2018). Kinetic modelling and simulation of batch, continuous and cell-recycling fermentations for acetone-butanol-ethanol production using clostridium saccharoperbutylacetonicum n1-4. Biochemical Engineering Journal, 137, 30–39.

Dochain, D. (2003). State and parameter estimation in chemical and biochemical processes: a tutorial. Journal of Process Control, 13(8), 801–818.

Doelle, H. W., Rokem, J. S., & Berovic, M. (2009). BIOTECHNOLOGY: Fundamentals in Biotechnology, volume 8. EOLSS Publications.

Donoso-Bravo, A., Pérez-Elvira, S., & Fdz-Polanco, F. (2014). Simplified mechanistic model for the two-stage anaerobic degradation of sewage sludge. Environmental Technology, 36(10), 1334–1346.

du Preez, R., Clarke, K. G., Callanan, L. H., & Burton, S. G. (2015). Modelling of immobilised enzyme biocatalytic membrane reactor performance. Journal of Molecular Catalysis B: Enzymatic, 119, 48–53.

Energy Insights - McKinsey (2019). Global Oil Supply and Demand Outlook. Technical report.

Englezos, P. & Kalogerakis, N. (2000). Applied parameter estimation for chemical engineers. CRC Press.

Farza, M., Hammouri, H., Othman, S., & Busawon, K. (1997). Nonlinear observers for parameter estimation in bioprocesses. Chemical Engineering Science, 52(23), 4251–4267.

Gadkar, K. G., Gunawan, R., & Doyle, F. J. (2005). Iterative approach to model identification of biological networks. BMC bioinformatics, 6(1), 1–20.

Gauthier, J. P., Hammouri, H., & Othman, S. (1992). A simple observer for nonlinear systems applications to bioreactors. IEEE Transactions on Automatic Control, 37(6), 875–880.

Gerland, P., Raftery, A. E., Ševčíková, H., Li, N., Gu, D., Spoorenberg, T., Alkema, L., Fosdick, B. K., Chunn, J., Lalic, N., et al. (2014). World population stabilization unlikely this century. Science, 346(6206), 234–237.

Grieg-Gran, M., Chomitz, K., Hyde, B., Muñoz, C., Richards, M., Sedjo, R., Stage, J., Steele, P., Tomich, T., Vargas, M.-t., et al. (2006). The cost of avoiding deforestation: Report prepared for the Stern Review of the economics of climate change. Technical report, International Institute for Environment and Development.

Grosfils, A., Wouwer, A. V., & Bogaerts, P. (2007). On a general model structure for macroscopic biological reaction rates. Journal of Biotechnology, 130(3), 253–264.

Henninger, H. B., Reese, S. P., Anderson, A. E., & Weiss, J. A. (2010). Validation of computational models in biomechanics. Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine, 224(7), 801–812.

Hernández-Escoto, H., Prado-Rubio, O. A., & Morales-Rodriguez, R. (2016). Model-based framework for enhanced and controlled operation of a fed-batch bioreactor: xylitol production. In Computer Aided Chemical Engineering, volume 38 (pp. 301–306). Elsevier.

Hernández-Pérez, A., Costa, I., Silva, D., Dussán, K., Villela, T., Canettieri, E., Carvalho Jr, J., Neto, T. S., & Felipe, M. (2016). Biochemical conversion of sugarcane straw hemicellulosic hydrolyzate supplemented with co-substrates for xylitol production. Bioresource technology, 200, 1085–1088.

Hulhoven, X., Wouwer, A. V., & Bogaerts, P. (2008). State observer scheme for joint kinetic parameter and state estimation. Chemical Engineering Science, 63(19), 4810–4819.

Illanes, A. & Wilson, L. (2003). Enzyme reactor design under thermal inactivation. Critical Reviews in Biotechnology, 23(1), 61–93.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 82(1), 35–45.

Keesman, K. J. (2011). System identification: an introduction. Springer Science & Business Media.

Kobayashi, F. & Nakamura, Y. (2004). Effect of repressor gene on stability of bioprocess with continuous conversion of starch into ethanol using recombinant yeast. Biochemical Engineering Journal, 18(2), 133–141.

Kolewe, M. E., Roberts, S. C., & Henson, M. A. (2011). A population balance equation model of aggregation dynamics in taxus suspension cell cultures. Biotechnology and Bioengineering, 109(2), 472–482.

Kosorok, M. R. (2008). Introduction to empirical processes. Springer.

Koza, J. R. (1990). Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems. Stanford University, Department of Computer Science Stanford, CA.

Kumar, V., Sandhu, P. P., Ahluwalia, V., Mishra, B. B., & Yadav, S. K. (2019). Improved upstream processing for detoxification and recovery of xylitol produced from corncob. Bioresource technology, 291, 121931.

Lema-Perez, L., Muñoz-Tamayo, R., Garcia-Tirado, J., & Alvarez, H. (2019). On parameter interpretability of phenomenological-based semiphysical models in biology. Informatics in Medicine Unlocked, 15, 100158.

Ling, Y. & Mahadevan, S. (2013). Quantitative model validation techniques: New insights. Reliability Engineering & System Safety, 111, 217–231.

Ljung, L. (1990). System Identification: Theory for the User. Prentice-Hall.

Ljung, L. & Glad, T. (1994). On global identifiability for arbitrary model parametrizations. Automatica, 30(2), 265–276.

Martínez, E. A., Canettieri, E. V., Bispo, J. A., Giulietti, M., De Almeida e Silva, J. B., & Converti, A. (2015). Strategies for xylitol purification and crystallization: a review. Separation Science and Technology, 50(14), 2087–2098.

Mašić, A., Srinivasan, S., Billeter, J., Bonvin, D., & Villez, K. (2017). Shape constrained splines as transparent black-box models for bioprocess modeling. Computers & Chemical Engineering, 99, 96–105.

McDonald, M. A., Bommarius, A. S., Rousseau, R. W., & Grover, M. A. (2019). Continuous reactive crystallization of β-lactam antibiotics catalyzed by penicillin g acylase. part i: Model development. Computers & Chemical Engineering, 123, 331–343.

Moreira, H., Pereira, S., & Castro, P. (2016). Impact of heavy metals and metalloids on soil microorganisms of mining areas, chapter 4, (pp. 95–120). Nova Science Publishers: .

Nadri, M., Trezzani, I., Hammouri, H., Dhurjati, P., Longin, R., & Lieto, J. (2006). Modeling and observer design for recombinant escherichia coli strain. Bioprocess and Biosystems Engineering, 28(4), 217–225.

Narayanan, H., Luna, M. F., von Stosch, M., Cruz Bournazou, M. N., Polotti, G., Morbidelli, M., Butté, A., & Sokolov, M. (2020). Bioprocessing in the digital age: the role of process models. Biotechnology journal, 15(1), 1900172.

National Research Council (2012). Assessing the Reliability and of Complex and Models: and Mathematical and and Statistical Foundations of Verification and Validation and Uncertainty Quantification. National Academies Press.

Nayak, R. & Gomes, J. (2009). Sequential adaptive networks: An ensemble of neural networks for feed forward control of l-methionine production. Chemical Engineering Science, 64(10), 2401–2412.

Neba, F. A., Asiedu, N. Y., Morken, J., Addo, A., & Seidu, R. (2020). A novel simulation model, bk_biogasim for design of onsite anaerobic digesters using two-stage biochemical kinetics: Codigestion of blackwater and organic waste. Scientific African, 7, e00233.

Organización de las Naciones Unidas (1992). Convenio de diversidad biológica. Technical report.

Pelletier, G. J., Chapra, S. C., & Tao, H. (2006). Qual2kw: A framework for modeling water quality in streams and rivers using a genetic algorithm for calibration. Environmental Modelling & Software, 21(3), 419–425.

Pfaffinger, C. E., Severin, T. S., Apel, A. C., Göbel, J., Sauter, J., & Weuster-Botz, D. (2019). Light-dependent growth kinetics enable scale-up of well-mixed phototrophic bioprocesses in different types of photobioreactors. Journal of biotechnology, 297, 41–48.

Prado-Rubio, O. A. (2010). Integration of Bioreactor and Membrane Separation Processes: A Model Based Approach: Reverse Electro-Enhanced Dialysis process for lactic acid fermentation. PhD thesis, Technical University of Denmark.

Prado-Rubio, O. A., Hernández-Escoto, H., Rodriguez-Gomez, D., Sirisansaneeyakul, S., & Morales-Rodriguez, R. (2015). Enhancing xylitol bio-production by an optimal feeding policy during fed-batch operation. In Computer Aided Chemical Engineering, volume 37 (pp. 1757–1762). Elsevier.

Rajamanickam, V., Babel, H., Montano-Herrera, L., Ehsani, A., Stiefel, F., Haider, S., Presser, B., & Knapp, B. (2021). About model validation in bioprocessing. Processes, 9(6), 961.

Rakestraw, J., Baskaran, A., & Wittrup, K. (2006). A flow cytometric assay for screening improved heterologous protein secretion in yeast. Biotechnology Progress, 22(4), 1200–1208.

Rana, K. L., Kour, D., Sheikh, I., Dhiman, A., Yadav, N., Yadav, A. N., Rastegari, A. A., Singh, K., & Saxena, A. K. (2019). Endophytic fungi: biodiversity, ecological significance, and potential industrial applications. In Recent advancement in white biotechnology through fungi (pp. 1–62). Springer.

Rehman, S., Nadeem, M., Ahmad, F., & Mushtaq, Z. (2018). Biotechnological production of xylitol from banana peel and its impact on physicochemical properties of rusks.

Research G. V. (2017). Biotechnology market analysis by application (health, food & agriculture, natural resources & environment, industrial processing bioinformatics), by technology, and segment forecasts, 2014 - 2025. Technical report.

Roman, M. & Selişteanu, D. (2012). Enzymatic synthesis of ampicillin: Nonlinear modeling, kinetics estimation, and adaptive control. Journal of Biomedicine and Biotechnology, 2012, 1–14.

Rosser, A. M. & Mainka, S. A. (2002). Overexploitation and species extinctions. Conservation Biology, 16(3), 584–586.

Sagmeister, P., Wechselberger, P., Jazini, M., Meitz, A., Langemann, T., & Herwig, C. (2013). Soft sensor assisted dynamic bioprocess control: Efficient tools for bioprocess development. Chemical Engineering Science, 96, 190–198.

Sai, Y., Siva Kishore, N., Dattatreya, A., Anand, S., & Sridhari, G. (2011). A review on biotechnology and its commercial and industrial applications. Journal of Biotechnology & Biomaterials, 01(07).

Sargent, R. G. (2013). Verification and validation of simulation models. Journal of simulation, 7(1), 12–24.

Sarmah, N., Revathi, D., Sheelu, G., Yamuna Rani, K., Sridhar, S., Mehtab, V., & Sumana, C. (2018). Recent advances on sources and industrial applications of lipases. Biotechnology progress, 34(1), 5–28.

Schoefs, O., Perrier, M., Dochain, D., & Samson, R. (2003). On-line estimation of biodegradation in an unsaturated soil. Bioprocess and Biosystems Engineering, 26(1), 37–48.

Seber, G. A. & Wild, C. J. (2003). Nonlinear Regression, volume 62. Wiley Series in Probability and Statistics.

Selişteanu, D., Tebbani, S., Roman, M., Petre, E., & Georgeanu, V. (2014). Microbial production of enzymes: Nonlinear state and kinetic reaction rates estimation. Biochemical Engineering Journal, 91, 23–36.

Shylesh, S., Gokhale, A. A., Ho, C. R., & Bell, A. T. (2017). Novel strategies for the production of fuels, lubricants, and chemicals from biomass. Accounts of chemical research, 50(10), 2589–2597.

Sonnleitnert, B., Kappeli, ., for Biotechnolog y, D., of Technology, S. F. I., Zurich, ., & Switzerland (1986). Growth of saccharomyces cerevisiae is controlled by its limited respiratory capacity: Formulation and verification of a hypothesis. Biotechnology and Bioengineering, 28(6), 927–937.

Spatz, D. R., Zilliacus, K. M., Holmes, N. D., Butchart, S. H., Genovesi, P., Ceballos, G., Tershy, B. R., & Croll, D. A. (2017). Globally threatened vertebrates on islands with invasive species. Science Advances, 3(10), e1603080.

Stigler, S. M. (1981). Gauss and the invention of least squares. The Annals of Statistics, 9(3), 465–474.

Sun, J., Smets, I., Bernaerts, K., Van Impe, J., Vanderleyden, J., & Marchal, K. (2001). Quantitative analysis of bacterial gene expression by using the gusa reporter gene system. Applied and environmental microbiology, 67(8), 3350–3357.

Surribas, A., Montesinos, J. L., & Valero, F. F. (2006). Biomass estimation using fluorescence measurements in pichia pastoris bioprocess. Journal of Chemical Technology and Biotechnology, 81(1), 23–28.

Szombatfalvy, L. (2010). The Greatest Challenges of Our Time. Ekerlids Publishing House.

Tajam, J., Mokhtar, M., & Wagiman, S. (2017). Mini review of coral reefs status in langkawi unesco global geopark (lugg), malaysia. Ecology, Environment and Conservation Paper, 23.

Tang, P., Xu, J., Louey, A., Tan, Z., Yongky, A., Liang, S., Li, Z. J., Weng, Y., & Liu, S. (2020). Kinetic modeling of chinese hamster ovary cell culture: factors and principles. Critical reviews in biotechnology, 40(2), 265–281.

Tasseff, R. & Varner, J. (2010). Mathematical models in biotechnology. PLoS One, 5, e8864.

Tochampa, W., Sirisansaneeyakul, S., Vanichsriratana, W., Srinophakun, P., Bakker, H. H., Wannawilai, S., & Chisti, Y. (2015). Optimal control of feeding in fed-batch production of xylitol. Industrial & Engineering Chemistry Research, 54(7), 1992–2000.

van Can, H. J. L., te Braake, H. A. B., Bijman, A., Hellinga, C., Luyben, K. C. A. M., & Heijnen, J. J. (1999). An efficient model development strategy for bioprocesses based on neural networks in macroscopic balances: Part ii. Biotechnology and Bioengineering, 62(6), 666–680.

Van den Bos, A. (2007). Parameter estimation for scientists and engineers. John Wiley & Sons.

Varga, E. G., Titchener-Hooker, N. J., & Dunnill, P. (2001). Prediction of the pilot-scale recovery of a recombinant yeast enzyme using integrated models. Biotechnology and bioengineering, 74(2), 96–107.

Vlassides, S., Ferrier, J. G., & Block, D. E. (2001). Using historical data for bioprocess optimization: Modeling wine characteristics using artificial neural networks and archived process information. Biotechnology and Bioengineering, 73(1), 55–68.

Wang, F.-S., Su, T.-L., & Jang, H.-J. (2001). Hybrid differential evolution for problems of kinetic parameter estimation and dynamic optimization of an ethanol fermentation process. Industrial & Engineering Chemistry Research, 40, 2876–2885.

Werpy, T. & Petersen, G. (2004). Top value added chemicals from biomass: volume I - Results of screening for potential candidates from sugars and synthesis gas. Technical report, Pacific Northwest National Laboratory (PNNL), National Renewable Energy Laboratory (NREL) & Office of Biomass Program (EERE).

Wu, Y., Lu, J., Sun, Y., & Yu, P. (2005). Bioprocess modeling using genetic programming based on a double penalty strategy. In International Conference on Computational and Information Science (pp. 921–926).: Springer.

Yeo, D., Kiparissides, A., Cha, J. M., Aguilar-Gallardo, C., Polak, J. M., Tsiridis, E., Pistikopoulos, E. N., & Mantalaris, A. (2013). Improving embryonic stem cell expansion through the combination of perfusion and bioprocess model design. PLoS ONE, 8(12), e81728.

Zabot, G. L., Mecca, J., Mesomo, M., Silva, M. F., Prá, V. D., de Oliveira, D., Oliveira, J. V., Castilhos, F., Treichel, H., & Mazutti, M. A. (2011). Hybrid modeling of xanthan gum bioproduction in batch bioreactor. Bioprocess and Biosystems Engineering, 34(8), 975–986.

Zhang, L., Song, Z., Pan, X., Feng, M., & Jin, Z. (2011). Comparison of v-support vector regression and logistic equation for descriptive modeling of lactobacillus plantarum growth. African Journal of Biotechnology, 10(32), 6162–6171.

Zhou, Q., Liu, Y., & Yuan, W. (2020). Kinetic modeling of butyric acid effects on butanol fermentation by clostridium saccharoperbutylacetonicum. New biotechnology, 55, 118–126.

Capítulo 2

Tratamiento de datos para estimación de parámetros

2.1. Resumen

La toma o recolección de datos es un paso fundamental en cualquier experimento, que se realiza con el fin de obtener información deseada sobre un sistema en específico. En el caso de los modelos matemáticos de procesos biotecnológicos, la información experimental toma gran relevancia, puesto que, a través de comparación, permite encontrar los parámetros desconocidos de un modelo. Sin embargo, los datos experimentales están sujetos a diferentes fuentes de error que alteran las mediciones y por tanto el valor observado. Los tipos de error pueden dividirse en aleatorios y sistemáticos, los primeros causados por eventos derivados de perturbaciones aleatorias en la medición. El segundo tipo de error implica la generación de tendencias que originalmente no existen en las mediciones. Puesto que el error intrínseco de los datos experimentales será propagado a los parámetros, y por tanto a las predicciones del modelo, se hace necesario verificar la calidad de los mismos. Entre ellas se tienen la detección de datos atípicos o "outliers" y el suavizado de datos. La detección de outliers corresponde a la identificación de valores que tienen una baja probabilidad de ocurrencia y pueden atribuirse a un error de medición. El suavizado de datos es un procedimiento que permite eliminar principalmente el ruido aleatorio presente en un conjunto de datos. Debido a que en general los modelos matemáticos de procesos biotecnológicos son de naturaleza determinística, eliminar el ruido en los datos experimentales mejora el desempeño del modelo. Lo anterior es posible a causa de que el modelo describirá de mejor manera el comportamiento del sistema en lugar de ruido experimental. En este capítulo se evaluará una metodología de limpieza de datos experimentales que consta en detección de puntos atípicos, su reemplazo y finalmente el suavizado de datos por aproximación polinomial.

2.2. Introducción

Los modelos matemáticos aplicados para la descripción de procesos biotecnológicos pueden ser creados en forma de caja negra, caja blanca o caja gris. Los modelos de caja negra son construidos a partir de datos experimentales, tomando en cuenta solo variables de entrada y salida del sistema. Los modelos de caja blanca se basan únicamente en conocimiento teórico del proceso, es decir, relaciones fenomenológicas y leyes físicas, por tanto, no requieren de estimación de parámetros. Finalmente, los modelos de caja gris corresponden a modelos mixtos construidos a partir de conocimientos teóricos referentes a la fenomenología del proceso y parámetros usualmente desconocidos cuyo valor debe estimarse de algún modo (Nelles, 2001).

La estimación de parámetros requiere entonces de *datos experimentales* obtenidos del sistema a modelar (Cameron & Hangos, 2001). La adquisición dicha información involucra experimentación, y por tanto, está sujeta a *error de medición* definido como "la diferencia entre el valor **verdadero** y el valor **medido**" de la variable (Kaloyerou, 2018). El error de medición puede dividirse en dos tipos:

- **Error aleatorio**: se produce de forma "natural" derivado de la imposibilidad de controlar absolutamente todas las condiciones que pueden presentarse en un experimento al momento de realizar la medición. Se manifiesta como medidas con valores al rededor del valor verdadero de la media para esa medida.

- **Error sistemático**: es consistente y se produce en todas las mediciones de una determinada variable. Este tipo de error está sujeto a diferentes causas físicas como desajuste en el instrumento de medida y ocasiona que la media de la medida se aleje de su valor verdadero.

Un tipo particular de error aleatorio que puede surgir durante el desarrollo de las mediciones en un experimento son los "*valores atípicos o outliers*". Estos errores son considerados como valores que se desvían significativamente de la tendencia de su vecindad de datos (Quinn & Keough, 2002). Técnicamente, un *outlier* es considerado como un valor que se encuentra entre 3 o 4 desviaciones estándar de la media del conjunto de datos (vecindad), cuya probabilidad de ocurrencia es inferior al 0.1 % (Englezos & Kalogerakis, 2000). El efecto de la presencia de valores atípicos durante la estimación de parámetros ya ha sido estudiado anteriormente, encontrándose que afecta el valor de los parámetros obtenidos (Chang et al., 1988; Chen & Liu, 1993; Chen et al., 2019).

En cuanto a los errores aleatorios "comunes" se han desarrollado diversas metodologías para su tratamiento las cuales surgen de la teoría de procesamiento de señales y reciben el nombre de "filtros" (Ljung, 1990). Un filtro corresponde a una formulación matemática que permite extraer información de un conjunto de datos (señal) que posee ciertas características, en

este caso particular, ruido aleatorio (Giron-Sierra, 2016). La aplicación de estos filtros a un conjunto de datos recibe el nombre de "suavizado de datos".

Un estudio realizado sobre el efecto del suavizado de datos experimentales en estimación de parámetros para el cálculo de curvas de lactancia multifásicas, arrojó que un suavizado global de datos mejora significativamente el valor de la función objetivo y los intervalos de confianza de los parámetros (Gipson et al., 1990). En un estudio más reciente, se integró el suavizado de datos con el proceso de estimación de parámetros a través de diferentes filtros de Kalman, en el cual se calculó de manera simultánea los estados de un sistema de fermentación batch y otro sistema con fermentación continua junto con los parámetros desconocidos. Se encontró que los parámetros convergen de una manera rápida a un valor cercano al verdadero (Chitralekha et al., 2010).

Otra aplicación del tratamiento de datos útil para la estimación de parámetros corresponde a la inferencia de puntos intermedios entre los puntos experimentales. Normalmente, esta aplicación se realiza con métodos de trazadores interpolantes y un caso concreto de aplicación fue la estimación de parámetros de un modelo de producción de anticuerpos monoclonales de células de mamífero (Selişteanu et al., 2015). Recientemente, se ha destacado la importancia del tratamiento de datos para la estimación de parámetros en modelos basados en red metabólica, los cuales involucran una alta cantidad de estados y de parámetros lo cual hace más complejo este procedimiento (Hirai & Shiraishi, 2018).

Debido a que el proceso de estimación de parámetros depende entonces de los datos experimentales, el error en los mismos debe ser analizado. Bajo este escenario, el procesamiento de datos puede convertirse en una valiosa herramienta para mejorar la estimación de parámetros al eliminar errores cuya incertidumbre puede propagarse a los parámetros de modelo. Lo anterior se vuelve especialmente relevante al considerar la naturaleza no convexa del problema de optimización para estimación de parámetros. Un problema no convexo puede presentar múltiples combinaciones de parámetros que se verían aumentadas debido a errores en los datos experimentales, lo cual en última instancia disminuye la certeza en el valor estimado de los parámetros.

Dado que los *outliers* son una fuente significativa de error, su eliminación puede mejorar sustancialmente el proceso de optimización, dada la naturaleza de los modelos basados en ecuaciones diferenciales. Estos modelos matemáticos poseen una única trayectoria que será determinada solo por el valor de los parámetros y las condiciones iniciales de los estados representados por el modelo (Jost, 2013). Así, con la eliminación de error aleatorio presente en los datos experimentales, es posible reducir la cantidad de trayectorias que puede adoptar la salida del modelo. Esto puede incrementar la precisión y exactitud de la estimación de parámetros (Aster et al., 2005).

En este capítulo se explorarán diferentes metodologías de tratamiento de datos con el objetivo de detectar, eliminar y reemplazar *outliers*, así como disminuir ruido aleatorio presente en datos experimentales. Diversas metodologías serán expuestas, junto a su aplicación a datos experimentales reales. El efecto del pretratamiento de datos sobre la estimación de parámetros será analizado en el capítulo 5 de esta tesis.

2.3. Metodología

El esquema general de limpieza de datos se muestra en la Figura **2-1**. Se parte de datos experimentales y se realizan dos procesos consecutivos: detección y reemplazo de outliers, y suavizado de datos. Esta metodología surge como resultado de la presente investigación.

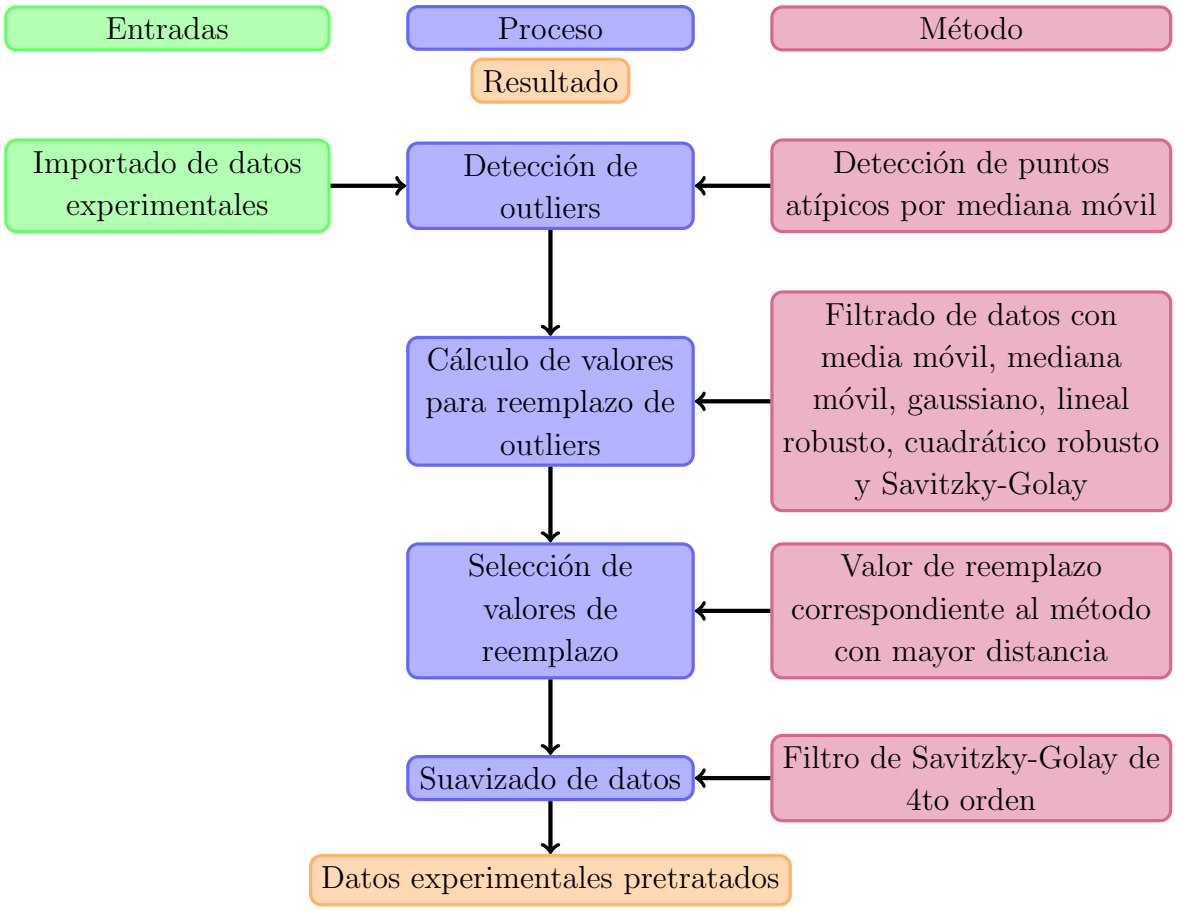


Figura 2-1: Metodología de limpieza de datos experimentales.

2.3.1. Detección de datos atípicos o *outliers*

La detección de *outliers* involucra la comparación del valor de cada uno de los datos de un conjunto respecto a una medida de tendencia central de ese conjunto. Si el valor de un dato cumple con algún criterio como ser mayor un número de desviaciones estándar de la media o la mediana se cataloga como un *outlier*. En Matlab® R2018b este procedimiento puede realizarse mediante el comando `isoutlier()`, el cual posee las siguientes opciones:

$$por\,media:\ outlier\ \geq 3\,\sigma \tag{2-1}$$

$$por\,mediana:\ outlier\ \geq 3\,cMDA \tag{2-2}$$

$$c = \left[\frac{-1}{\sqrt{2}\,\Gamma^{-1}(3/2)}\right] \tag{2-3}$$

$$MDA = mediana\,|A_i - mediana(A)| \tag{2-4}$$

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt \tag{2-5}$$

en donde σ corresponde a la desviación estándar del conjunto de datos A, c coeficiente de corrección probabilista, MDA desviación mediana absoluta, Γ^{-1} función inversa de la función de error (Γ) y z es la probabilidad que el dato sea un *outlier*. La primera opción del comando `isoutlier()` se considera como la definición "clásica" de outlier, sin embargo, posee baja sensibilidad en conjuntos de datos que son aparentemente homogéneos. Para el caso de la detección por mediana, esta se encuentra modificada en Matlab® utilizando el término c cuyo propósito es brindar una corrección estadística con un margen de probabilidad del 95 % para la mediana, lo que otorga mayor sensibilidad en la detección (Ruppert, 2011).

Como caso particular para los procesos biotecnológicos, se tienen datos de naturaleza dinámica, por tanto, se debe modificar el método convencional para escoger una vecindad de datos en la cual se realizará el análisis de detección. Lo anterior implica el uso de una ventana móvil. La opción correspondiente en Matlab® es `isoutlier(A,'movmean',R)` en caso de que se use la media y `isoutlier(A,'movmedian',R)` en caso de la mediana. El valor R es el número de puntos de la ventana móvil. La Figura **2-2** muestra la aplicación del comando `isoutlier(A,'movmedian',5)` (MatLab, 2018).

2.3.2. Reemplazo de puntos atípicos o *outliers*

Los *outliers* encontrados pueden ser reemplazados con un valor que se aproxime a la tendencia de la vecindad de datos en la que se encuentra dicho punto, esto con el fin de no perder por completo la información aportada por el dato atípico. Existen diferentes aproximaciones para analizar la vecindad de datos, y por ende, diferentes maneras de calcular el valor de reemplazo

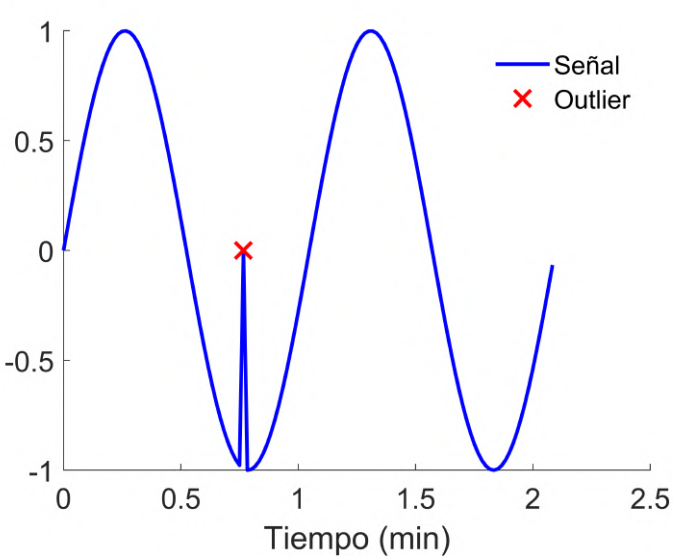


Figura 2-2: Ejemplo de outlier detectado en datos de una señal senosoidal.

para el *outlier*. Diferentes filtros de señal han sido utilizados para la eliminación de ruido aleatorio, recuperando la forma original de la señal. Los filtros se basan en una formulación matemática que transforma datos de entrada en datos de salida con alguna característica (Giron-Sierra, 2016). En este caso, los filtros son utilizados para calcular el posible valor que tendría el dato catalogado como *outlier* sí la medición hubiera sido realizada correctamente. El comando `smoothdata()` perteneciente al *signal processing toolbox* de Matlab® R2018b, posee seis diferentes opciones de filtros de datos que corresponden a:

- *movmean*: aplica una media móvil sobre la ventana de datos seleccionada, ecuación 2-1 (Quinn & Keough, 2002).

- *movmedian*: aplica una mediana móvil sobre la ventana de datos seleccionada, ecuación 2-2 (Quinn & Keough, 2002).

- *gaussian*: aplica un promedio ponderado gaussiano sobre la ventana de datos seleccionada, sigue la fórmula (Huet et al., 2006):

$$y = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \tag{2-6}$$

  en donde y es el dato a reemplazar, x el conjunto de datos de la ventana móvil y σ la desviación estándar de x.

- *rlowess*: aplica una regresión lineal robusta sobre la ventana de datos seleccionada (Batmend & Perdukova, 2013).

$$y = \beta_0 + \beta_1 x;\ \beta_1 = \frac{n\sum_{i=1}^{n} t_i x_i - \sum_{i=1}^{n} t_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} t_i^2 - \left(\sum_{i=1}^{n} t_i\right)^2}; \qquad \beta_0 = \frac{n\sum_{i=1}^{n} t_i^2 - \beta_1 \sum_{i=1}^{n} t_i}{n} \tag{2-7}$$

en donde y es el dato a reemplazar, x el conjunto de datos de la ventana móvil, t valores de tiempo correspondientes a los datos de la ventana móvil y n tamaño de la ventana.

- *rloess*: aplica una regresión cuadrática robusta sobre la ventana de datos seleccionada (Huet et al., 2006):

$$y = \beta_0 + \beta_1 x + \beta_2 x^2;$$

$$\beta_1 = \frac{\left[\sum tx - \frac{\sum t \sum x}{n}\right] * \left[\sum t^4 - \frac{(\sum t^2)^2}{n}\right] - \left[\sum t^2 x - \frac{\sum t^2 \sum x}{n}\right] * \left[\sum t^3 - \frac{\sum t^2 \sum t}{n}\right]}{\left[\sum t^2 - \frac{(\sum t)^2}{n}\right] * \left[\sum t^4 - \frac{(\sum t^2)^2}{n}\right] - \left[\sum t^3 - \frac{\sum t^2 \sum t}{n}\right]^2}$$

$$\beta_2 = \frac{\left[\sum t^2 - \frac{(\sum t)^2}{n}\right] * \left[\sum t^2 x - \frac{\sum t^2 \sum x}{n}\right] - \left[\sum t^3 - \frac{\sum t^2 \sum t}{n}\right] * \left[\sum tx - \frac{\sum t \sum x}{n}\right]}{\left[\sum t^2 - \frac{(\sum t)^2}{n}\right] * \left[\sum t^4 - \frac{(\sum t^2)^2}{n}\right] - \left[\sum t^3 - \frac{\sum t^2 \sum t}{n}\right]^2}$$

$$\beta_0 = \frac{\sum x - \beta_1 \sum t - \beta_2 \sum t^2}{n} \tag{2-8}$$

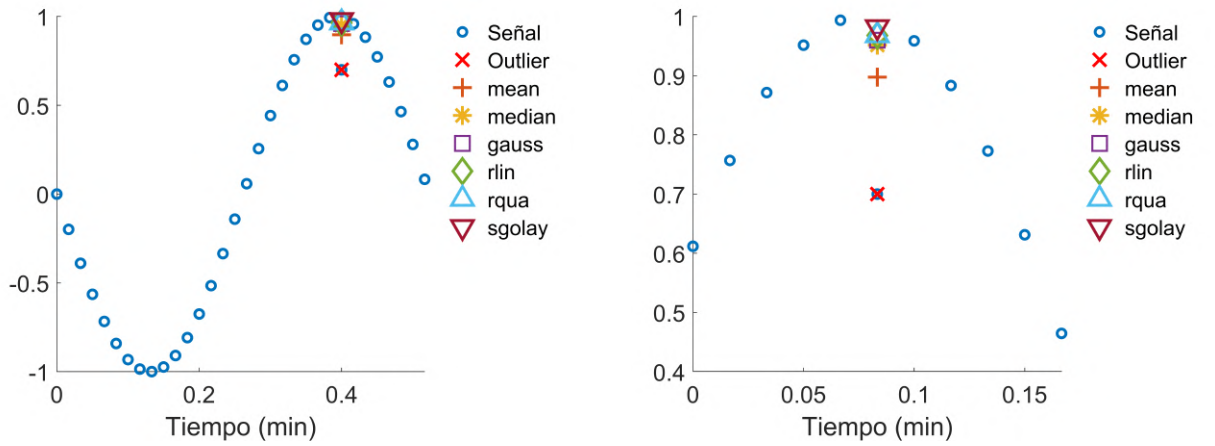
en donde y es el dato a reemplazar, x el conjunto de datos de la ventana móvil, t valores de tiempo correspondientes a los datos de la ventana móvil, n tamaño de la ventana y β_i coeficientes del polinomio cuadrático.

- *sgolay*: aplica el filtro de Saviztky-Golay, el cual genera un polinomio cuadrático ajustado sobre la ventana de datos seleccionada. Para una ventana de 5 puntos se tiene:

$$y_j = \frac{1}{35}(-3x_{j-2} + 12x_{j-1} + 17x_j + 12x_{j+1} - 3x_{j+2}) \tag{2-9}$$

en donde j hace referencia al punto a suavizar (Savitzky & Golay, 1964).

En la Figura **2-3**, se puede observar la detección de un *outlier* presente en una señal senosoidal y los valores de reemplazo para el punto atípico con los diferentes filtros del comando `smoothdata()`. Dado que los filtros poseen diferentes maneras de ajustar la tendencia general de la vecindad de datos y *a priori* no se puede conocer cuál va a generar un mejor ajuste, es recomendable emplear todos los filtros y reemplazar el *outlier* con el valor generado más lejano respecto al punto atípico.



(a) Señal con outlier y valores de reemplazo.

(b) Acercamiento a valores de reemplazo.

Figura 2-3: Detección y reemplazo de puntos atípicos. La notación de los filtros corresponde a: mean: media, median: mediana, gauss: gaussiano, rlin: lineal robusto, rqua: cuadrático robusto, sgolay: Savitzky-Golay.

2.3.3. Suavizado de datos por aproximación polinomial

El suavizado de datos es un procedimiento matemático que permite corregir distorsiones en la tendencia de un conjunto de datos o de una señal, esto es, eliminar ruido aleatorio presente en dichos datos (Giron-Sierra, 2016). Debido a que los datos experimentales utilizados para la estimación de parámetros contienen en mayor o menor grado ruido aleatorio, este procedimiento se hace necesario. Para esta investigación se utiliza un procedimiento de suavizado por aproximación polinomial que consiste en la implementación de un filtro de Savitzky-Golay de ventana móvil con un polinomio de cuarto orden (Savitzky & Golay, 1964). A continuación, en la Figura **2-4** se muestra un ejemplo propio de la aplicación de esta técnica sobre un conjunto de datos generados en Matlab® con la función seno y ruido aleatorio.

2.3.4. Indicador de desempeño

Una forma de determinar la cantidad de error aleatorio presente en los datos experimentales es la diferencia de estos sin tratar respecto a sí mismos una vez tratados. Un índice que puede aplicarse es el error porcentual absoluto medio (MAPE por sus siglas en inglés) descrito por la Fórmula 2-10, en donde y_i es el dato "crudo" y $y_{i,trat}$ es el dato tratado.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y_{i,trat}}{y_i}\right| \tag{2-10}$$

Este indicador describe el porcentaje total de desviación de los datos tratados respecto a los datos experimentales sin tratar. Una mayor desviación indicaría un mayor nivel de ruido aleatorio en los datos experimentales.

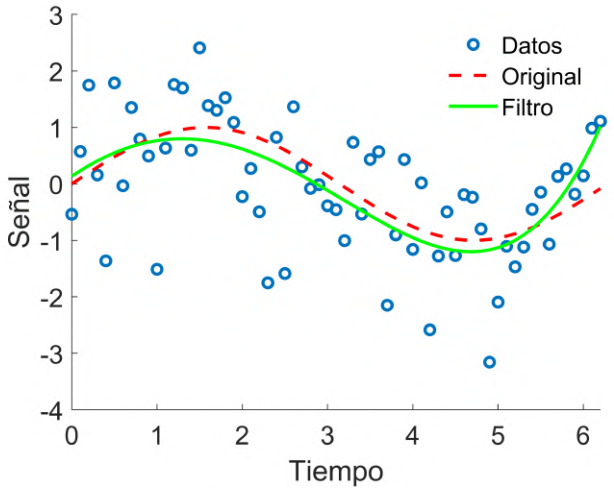


Figura 2-4: Suavizado de datos con ruido aleatorio.

2.4. Resultados

A continuación, se presentan los resultados de la aplicación de la metodología presentada a los conjuntos de datos experimentales 1 a 22 obtenidos de (Sirisansaneeyakul et al., 2013).

2.4.1. Detección y reemplazo de outliers en datos experimentales de fermentación diaúxica de glucosa y xilosa

Los conjuntos o "sets" de datos experimentales para la fermentación diaúxica de glucosa y xilosa fueron obtenidos de (Sirisansaneeyakul et al., 2013). En total se realizaron 22 experimentos de tipo *batch repetido*, con la característica que al finalizar una fermentación, sustrato adicional era agregado al biorreactor. El primer experimento fue realizado con ambos sustratos, los experimentos posteriores solo contaron con la adición de xilosa. La Figura **2-5** tomada del artículo muestra el comportamiento de las concentraciones de biomasa, glucosa, xilosa y xilitol en el transcurso de los experimentos.

La Figura **2-6** muestra el resultado de la detección y reemplazo de un *outlier* en la variable biomasa para el primer set de datos experimentales. El punto atípico presente en el tercer dato del conjunto de datos fue detectado mediante la opción `'movmedian'`, en contraste, fue considerado como un valor normal cuando se utiliza la opción `'movmean'`. Lo anterior exhibe la sensibilidad de las Fórmulas 2-2, por lo cual esta opción se seguirá utilizando en esta investigación para la detección de *outliers*. El valor de reemplazo para el outlier fue calculado con las opciones antes mencionadas del comando `smoothdata()`, y escogido mediante el criterio de máxima distancia respecto al *outlier*.

En términos generales, la variable con mayor cantidad de *outliers* es la concentración de bio-

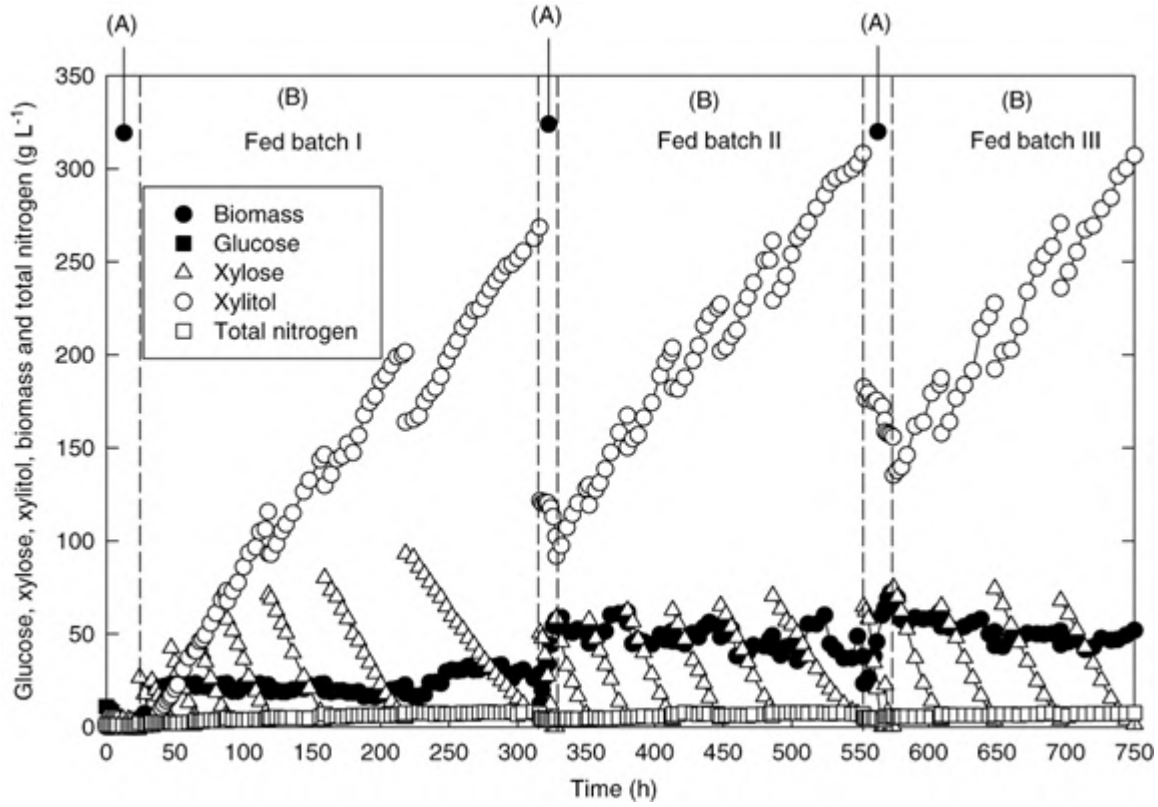
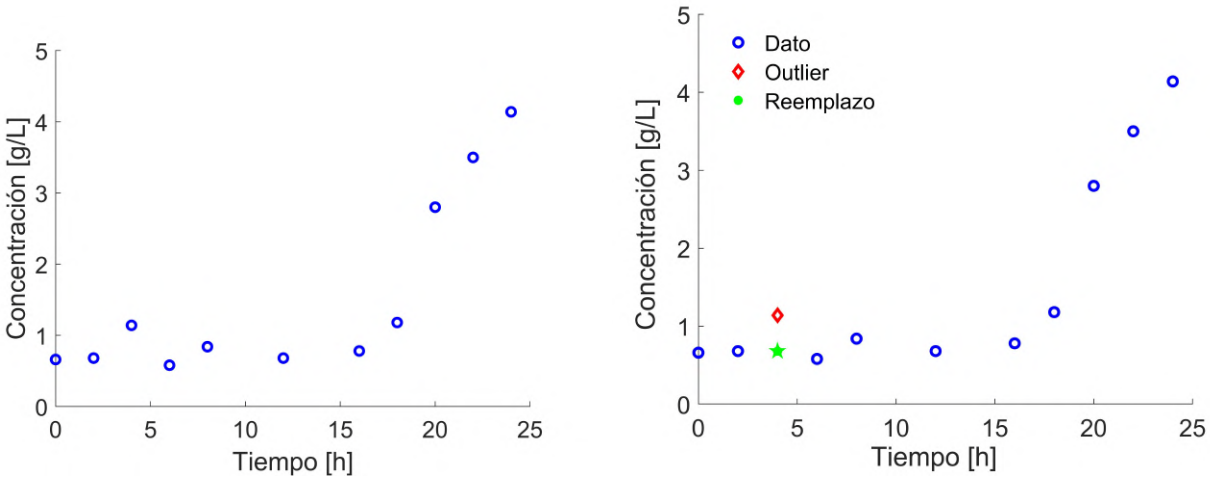


Figura 2-5: Fermentaciones en modalidad de batch repetidos tomada de (Sirisansaneeyakul et al., 2013). (A) experimento para crecimiento de biomasa, (B) experimento de batch repetido.

masa, esto debido a la técnica de muestreo por gravimetría. Esta técnica analítica presenta diferentes fuentes de error como resolución de la medición, ubicación y periodo de muestreo en biorreactor, rehidratación de la muestra, metabolismo celular activo o pérdida de células en el proceso de lavado (Sonnleitner et al., 1992; Arnáiz Franco et al., 2000). En contraste, las concentraciones de glucosa, xilosa y xilitol presentaron ruido aleatorio en lugar de *outliers*, lo que puede ser atribuido tanto a su cuantificación por colorimetría y al ser sustancias solubles se encuentran distribuidas de manera homogénea en el biorreactor durante la fermentación (Hollatz & Stambuk, 2001; Lai et al., 2016).

El procedimiento utilizado es entonces sencillo y eficiente para detectar y reemplazar *outliers* en datos de tipo *off-line*, lo cual mejora la calidad de los datos y evita la pérdida de la totalidad de información contenida en los puntos atípicos. Otras formulaciones para la detección de *outliers* han sido reportadas en literatura como el filtro revisado de Martin y Thompson (Liu et al., 2004) o el filtro de Hampfel (Pearson, 2002) que aunque son efectivos, requieren de implementaciones matemáticas complejas. La detección de *outliers* fue realizada de manera conjunta con el proceso de suavizado para los conjuntos experimentales 2 a 22. En términos

generales, la variable concentración de biomasa presento 12 puntos atípicos, la mayor cantidad en todos los conjuntos de datos, seguida de la variable xilitol con 5 y finalmente, xilosa con 1 *outlier*.



(a) Datos originales

(b) Detección y reemplazo de outlier

Figura 2-6: Detección de puntos atípicos en concentración de biomasa para set de datos 1.

2.4.2. Suavizado de datos de fermentación diaúxica de glucosa y xilosa

Los datos experimentales fueron tratados según el procedimiento mostrado en la Figura **2-1**. Los resultados de la limpieza de datos se muestra en las Figuras **2-7** a **2-10** para diferentes conjuntos de datos experimentales, específicamente los experimentos 1, 9, 10 y 22. Estos conjuntos de datos fueron seleccionados para su exhibición dado que hay diferencias notables entre los datos originales y los tratados. Una vez los *outliers* son removidos es necesario eliminar en la medida de lo posible el error aleatorio presente en los datos, esto con el fin de capturar netamente el comportamiento del sistema analizado. Los resultados de los conjuntos de datos restantes son mostrados en las Figuras **A-1** a **A-18**.

2.4.3. Evaluación del proceso de detección

La Tabla **2-1** presenta los índices MAPE calculados para los estados en los 22 conjuntos de datos experimentales analizados. Se observan valores MAPE superiores al 100 % en los conjuntos de datos 1, 9, 10, 16, 17 y 22. Entre ellos, se destacan los conjuntos 10 y 22 los cuales presentan valores MAPE superiores al 1000 %. Un valor MAPE superior a 100 % puede ser debido a la presencia de *outliers* dado que estos presentan una desviación anormal del valor que cabria esperar en su vecindad, situación que se presenta para el estado biomasa en

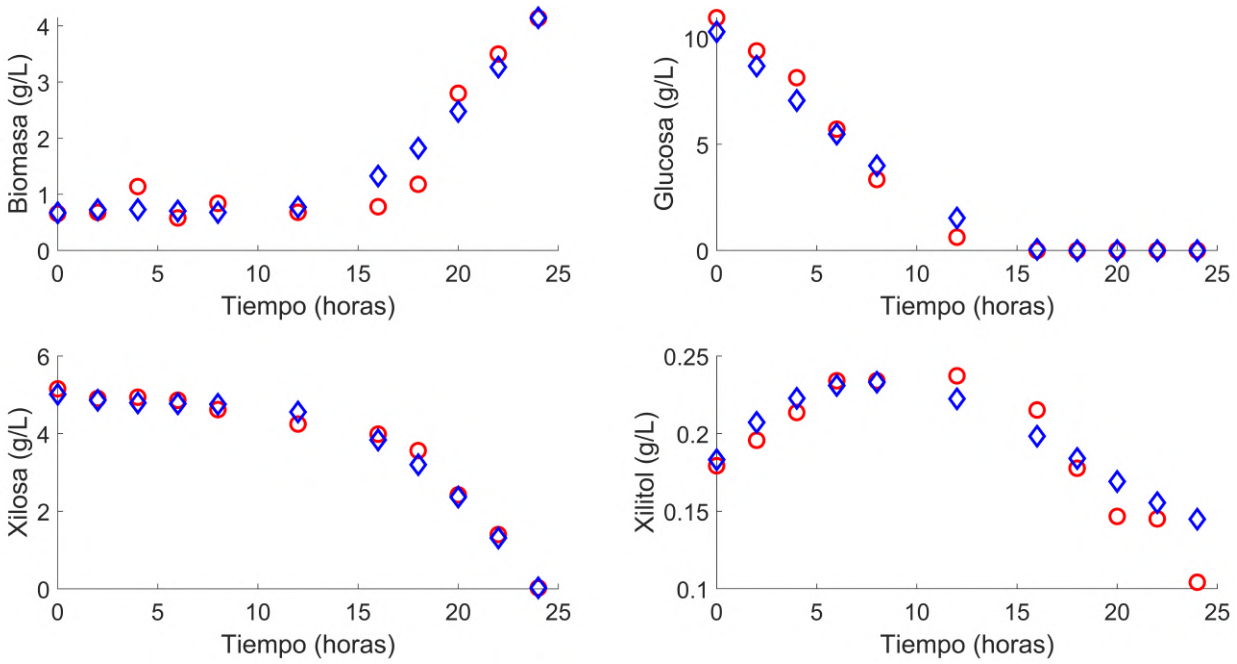


Figura 2-7: Conjunto de datos 1: (○) dato experimental, (◇) dato tratado.

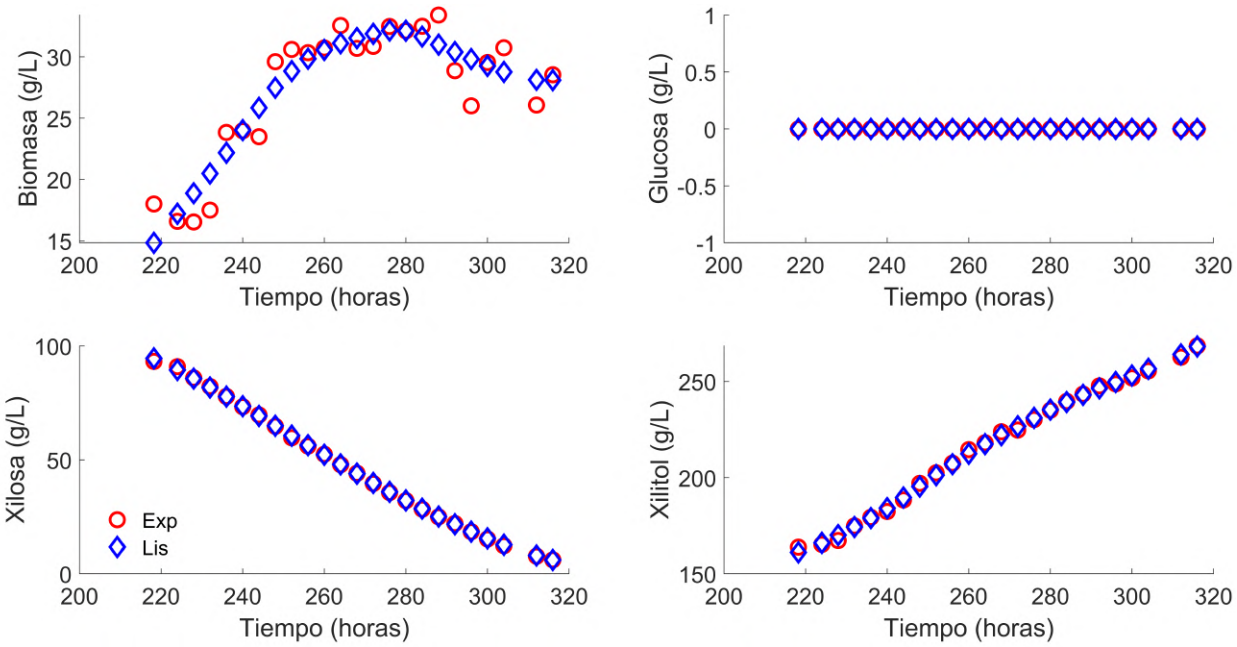


Figura 2-8: Conjunto de datos 9: (○) dato experimental, (◇) dato tratado.

los conjunto de datos 9 (Figura **2-8**) y 16 (ver Figura **A-13**). Valores MAPE superiores al 1000 % pueden ser explicados por la formación de una nueva tendencia en los datos pretratados, debido al proceso de suavizado. Esta situación se observa en el estado xilosa para los conjuntos de datos 10 (Figura **2-9**) y 17 (ver Figura **A-14**). Por otro lado, valores MAPE

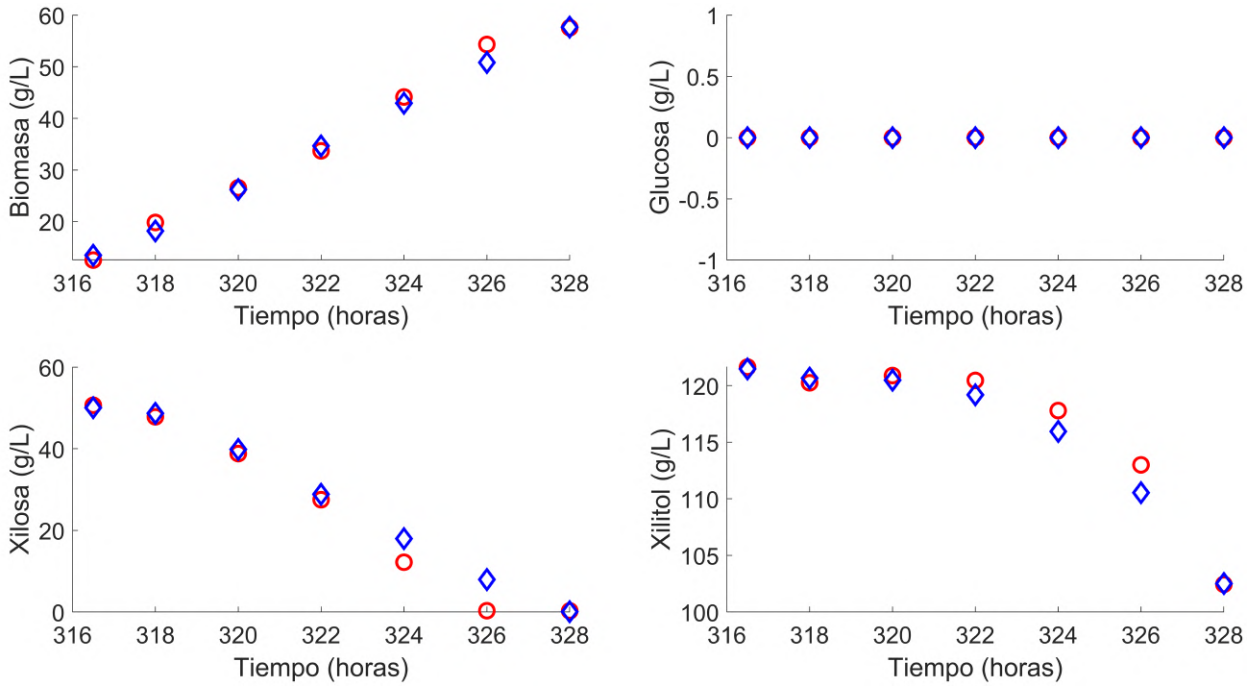


Figura 2-9: Conjunto de datos 10: (○) dato experimental, (◇) dato tratado.

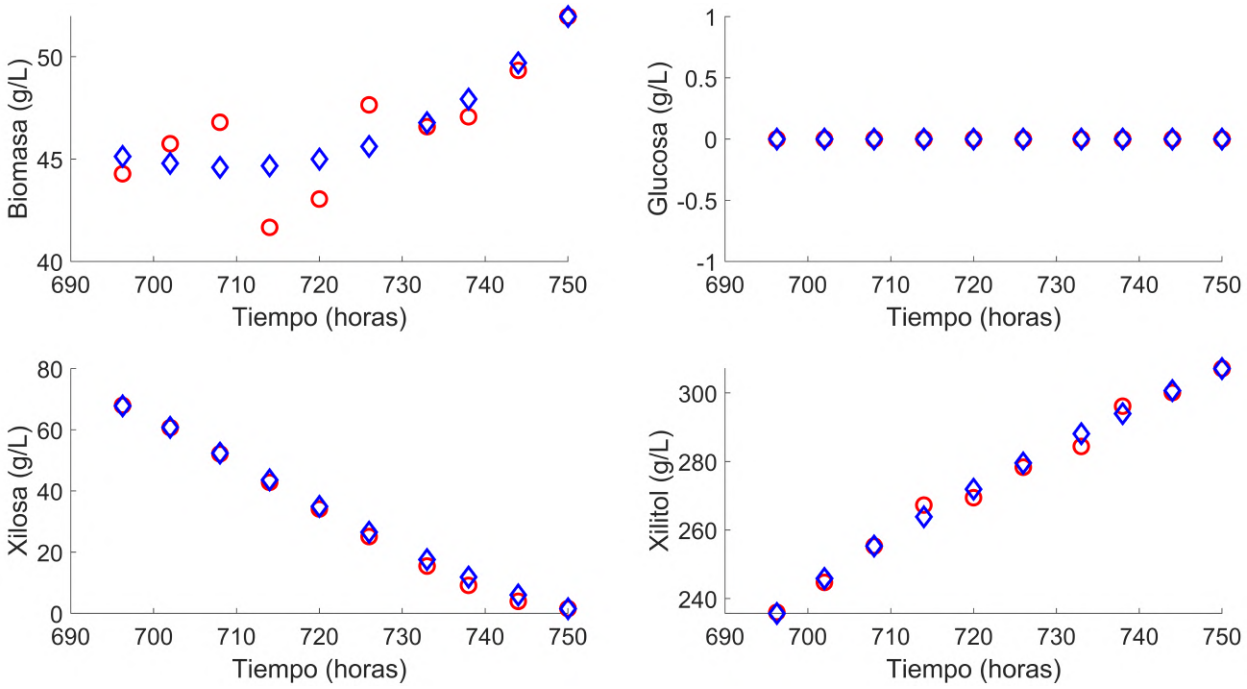


Figura 2-10: Conjunto de datos 22: (○) dato experimental, (◇) dato tratado.

inferiores o cercanos al 70 % indicarían que los datos experimentales presentan solamente ruido aleatorio como se observa en el estado xilitol para el conjunto de datos 9 (Figura **2-8**).

Tabla 2-1: Valores de error absoluto porcentual medio por variable y conjunto de datos.

Conjunto de datos	Variables			
	Biomasa	Glucosa	Xilosa	Xilitol
1	**242.19**	**200.52**	57.96	93.02
2	0.00	0.00	0.00	0.00
3	5.41	0.00	8.09	21.58
4	8.68	0.00	8.09	21.58
5	7.26	0.00	38.89	13.55
6	27.06	0.00	17.72	15.57
7	14.09	0.00	21.94	6.88
8	48.73	0.00	28.25	21.95
9	**142.71**	0.00	24.82	13.96
10	29.23	0.00	**2679.9**	5.68
11	21.79	0.00	11.23	3.72
12	18.68	0.00	10.34	3.24
13	32.93	0.00	28.21	6.93
14	28.23	0.00	28.71	10.03
15	68.12	0.00	16.14	9.77
16	**167.67**	0.00	39.20	9.27
17	21.48	0.00	**2291.7**	8.50
18	0.00	0.00	0.00	0.00
19	31.86	0.00	30.60	9.95
20	20.94	0.00	26.33	8.58
21	16.97	0.00	32.40	14.65
22	27.71	0.00	**109.97**	5.48

2.4.4. Directrices para pretratamiento de datos

El pretratamiento de datos entendido como la limpieza de puntos atípicos (outliers) y ruido aleatorio mejora el proceso de estimación de parámetros en modelos matemáticos basados en ecuaciones diferenciales principalmente por dos motivos: primero, la eliminación de puntos atípicos reduce la posibilidad de una desviación significativa de la trayectoria del modelo respecto a la trayectoria determinística de los datos experimentales. Adicionalmente, el reemplazo de los puntos atípicos permite conservar información experimental (relevante cuando se tiene poca cantidad de experimentos). Segundo, la reducción en el ruido aleatorio de los datos experimentales limita la cantidad de trayectorias estadísticamente posibles que puede adoptar el modelo matemático. Lo anterior es relevante al considerar la condición de Lipschitz (continuidad) de la cual se deriva la condición de unicidad de la solución de sistemas de ecuaciones diferenciales ordinarias, la cual establece que para un modelo, parámetros y condiciones iniciales definidas solo existe una única trayectoria (Bruckner et al., 2001; Soh-

rab, 2003). De esta manera, el pretratamiento de datos facilitaría la búsqueda del valor de los parámetros y simultáneamente, reduciría la propagación de incertidumbre experimental a los parámetros y predicción del modelo. Con base en la Figura **2-1**, las directrices de pretratamiento de datos sugeridas en esta tesis son:

1. Detección de puntos atípicos (outliers) mediante el método de mediana móvil: la mediana descrita por la Fórmula 2-2es una medida estadísticamente robusta y altamente sensible, la cual aplicada a una ventana de datos móvil permite la detección de outliers en conjuntos de datos dinámicos.

2. Cálculo de valores de reemplazo: para evitar pérdida de información experimental representada por el punto atípico primero se realiza un ajuste de datos con filtros de señal (media móvil, mediana móvil, gaussiano, lineal robusto, cuadrático robusto y filtro de Savitzky-Golay). El valor correspondiente al outilier es calculado con el ajuste obtenido para cada uno de los filtros de señal ya mencionados. El valor de reemplazo será el valor calculado que presente la mayor diferencia absoluta con respecto al punto atípico.

3. Suavizado de datos: los datos experimentales sin outliers serán filtrados con un polinomio de Savitzky-Golay de 4to orden. Este filtro de señal permite eliminar el ruido aleatorio de los datos experimentales mediante el ajuste a un polinomio el cual puede ser interpretado como la tendencia determinística del conjunto de datos (se hace referencia a la Figura **2-4**).

4. Validación de pretratamiento de datos: la calidad de los datos experimentales y el éxito del pretratamiento se cuantifican con el valor del índice MAPE (Fórmula 2-10) de la siguiente manera:

    4.1. Valor MAPE $< 70\,\%$: los datos experimentales presentan solamente ruido aleatorio.

    4.2. Valor MAPE $\geq 100\,\%$: los datos experimentales presentan por lo menos 1 punto atípico.

    4.3. Valor MAPE $\geq 900\,\%$: el suavizado de datos calculó una tendencia diferente a los datos experimentales. La nueva tendencia es producto de un artefacto del método, por tanto, se hace necesario modificar los parámetros del filtro de Savitzky-Golay y realizar una nueva validación.

2.5. Conclusiones

El tratamiento de datos es un procedimiento que reúne un conjunto de herramientas que buscan mejorar la calidad de los datos experimentales que, en este caso, van a ser utilizados

tanto para estimación como validación de parámetros. En este sentido, se pretende comprobar en capítulos posteriores el efecto de la eliminación de *outliers* junto con su reemplazo y el suavizado del conjunto de datos en la certeza de los parámetros estimados. Así mismo, el índice MAPE es útil para establecer, de manera cualitativa, la calidad inicial de los datos experimentales, en donde valores inferiores al 70 % indican un nivel de ruidos aleatorio razonable en los datos experimentales, valores alrededor del 100 % o superiores indican la presencia de *outliers* y valores muy superiores al 100 % indican una nueva tendencia en los datos suavizados respecto a los datos originales.

Bibliografía

Arnáiz Franco, C., Isac Oria, L., & Lebrato Martínez, J. (2000). Determinación de la biomasa en procesos biológicos. i métodos directos e indirectos. Tecnología del agua, 20 (205), 45-52.

Aster, R. C., Borchers, B., & Thurber, C. H. (2005). Parameter estimation and inverse problems. Elsevier.

Batmend, M. & Perdukova, D. (2013). Linear regression based real-time filtering. Advances in Electrical and Electronic Engineering, 11(6), 487–493.

Bruckner, A. M., Bruckner, J. B., & Thomson, B. S. (2001). Elementary real analysis. Prentice Hall.

Cameron, I. T. & Hangos, K., Eds. (2001). Process Modelling and Model Analysis, volume 4. Academic press.

Chang, I., Tiao, G. C., & Chen, C. (1988). Estimation of time series parameters in the presence of outliers. Technometrics, 30(2), 193–204.

Chen, B., Xing, L., Zhao, H., Du, S., & Príncipe, J. C. (2019). Effects of outliers on the maximum correntropy estimation: A robustness analysis. IEEE Transactions on Systems, Man, and Cybernetics: Systems.

Chen, C. & Liu, L.-M. (1993). Joint estimation of model parameters and outlier effects in time series. Journal of the American Statistical Association, 88(421), 284–297.

Chitralekha, S. B., Prakash, J., Raghavan, H., Gopaluni, R., & Shah, S. L. (2010). A comparison of simultaneous state and parameter estimation schemes for a continuous fermentor reactor. Journal of Process Control, 20(8), 934–943.

Englezos, P. & Kalogerakis, N. (2000). Applied parameter estimation for chemical engineers. CRC Press.

Gipson, T., Fernando, R., & Grossman, M. (1990). Effects of smoothing data on estimation of parameters for multiphasic lactation curves of dairy goats. Livestock Production Science, 24(3), 205–221.

Giron-Sierra, J. M. (2016). Digital Signal Processing with Matlab Examples, Volume 1: Signals and Data, Filtering, Non-stationary Signals, Modulation. Springer.

Hirai, M. Y. & Shiraishi, F. (2018). Using metabolome data for mathematical modeling of plant metabolic systems. Current opinion in biotechnology, 54, 138–144.

Hollatz, C. & Stambuk, B. U. (2001). Colorimetric determination of active α-glucoside transport in saccharomyces cerevisiae. Journal of microbiological methods, 46(3), 253–259.

Huet, S., Bouvier, A., Poursat, M.-A., & Jolivet, E. (2006). Statistical tools for nonlinear regression: a practical guide with S-PLUS and R examples. Springer Science & Business Media.

Jost, J. (2013). Partial Differential Equations. Graduate Texts in Mathematics. Springer, 3 edition.

Kaloyerou, P. N. (2018). Basic Concepts of Data and Error Analysis. Springer.

Lai, B., Plan, M. R., Hodson, M. P., & Krömer, J. O. (2016). Simultaneous determination of sugars, carboxylates, alcohols and aldehydes from fermentations by high performance liquid chromatography. Fermentation, 2(1), 6.

Liu, H., Shah, S., & Jiang, W. (2004). On-line outlier detection and data cleaning. Computers & chemical engineering, 28(9), 1635–1647.

Ljung, L. (1990). System Identification: Theory for the User. Prentice-Hall.

MatLab (2018). isoutlier() command documentation. MathWorks.

Nelles, O. (2001). Nonlinear system identification: from classical approaches to neural networks and fuzzy models. Springer Science & Business Media.

Pearson, R. K. (2002). Outliers in process modeling and identification. IEEE Transactions on control systems technology, 10(1), 55–63.

Quinn, G. P. & Keough, M. J. (2002). Experimental design and data analysis for biologists. Cambridge university press.

Ruppert, D. (2011). Statistics and data analysis for financial engineering, volume 13. Springer.

Savitzky, A. & Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. Analytical chemistry, 36(8), 1627–1639.

Selişteanu, D., Sendrescu, D., Georgeanu, V., & Roman, M. (2015). Mammalian cell culture process for monoclonal antibody production: nonlinear modelling and parameter estimation. BioMed research international, 2015.

Sirisansaneeyakul, S., Wannawilai, S., & Chisti, Y. (2013). Repeated fed-batch production of xylitol by candida magnoliae tistr 5663. Journal of Chemical Technology & Biotechnology, 88(6), 1121–1129.

Sohrab, H. H. (2003). Basic real analysis, volume 231. Springer.

Sonnleitner, B., Locher, G., & Fiechter, A. (1992). Biomass determination. Journal of Biotechnology, 25(1-2), 5–22.

Capítulo 3

Identificabilidad estructural de parámetros

3.1. Resumen

La identificabilidad estructural de parámetros corresponde a un análisis matemático que considera los estados, parámetros y estructura (forma de las ecuaciones) del modelo junto con los estados observados experimentalmente para determinar sí los parámetros estimados pueden ser conocidos a partir de la información disponible o no. Dado que los modelos matemáticos de caja gris buscan no solo representar un sistema, sino también brindar un entendimiento teórico del mismo cada parámetro posee un significado físico, es decir, posee interpretabilidad. Debido a lo anterior, se requiere certeza de que el valor numérico asignado al parámetro pueda conocerse y por tanto brindar información acerca del sistema modelado. En este capítulo se exponen algunas metodologías de análisis de identificabilidad estructural como lo son generación de series, álgebra diferencial y geometría diferencial, algunas herramientas informáticas basadas en estas metodologías y su aplicación al modelo. La identificabilidad estructural es un paso de validación fundamental para el modelo matemático pues sí los parámetros no son identificables estructuralmente, su valor numérico no puede ser estimado. Interesantemente, el análisis de identificabilidad estructural comúnmente no es realizado. La contribución "Análisis preliminares e identificabilidad práctica en modelos de procesos biotecnológicos" publicada en las memorias del evento AMIDIQ 2019 presenta los resultados obtenidos en este capítulo.

3.2. Introducción

En el capítulo anterior se describió como tratar los datos experimentales para reducir, en la medida de lo posible, error aleatorio y *outliers* que pueden representar un serio inconveniente al momento de estimar el valor del parámetro. En este capítulo se exhibirán estrategias de

identificabilidad estructural de parámetros y su aplicación al modelo matemático de fermentación diaúxica para producción de xilitol. La identificación práctica de parámetros corresponde al proceso de conocer el valor numérico de los parámetros de un modelo matemático, en este caso, de modelos matemáticos de procesos biotecnológicos (Englezos & Kalogerakis, 2000). Sin embargo, la estimación de parámetros presenta diferentes retos como error en las observaciones experimentales y diferencias de precisión entre variables y conjuntos de datos e incapacidad usual de cuantificar todos los estados que describe el modelo. Por otra parte, se puede presentar incertidumbre en los elementos que definen el problema de optimización, tales como algoritmos de optimización, función objetivo, entre otros (Biegler, 2010).

Es especialmente importante considerar el nivel de conocimiento disponible sobre el sistema que se pretende representar a través del modelo, puesto que a mayor nivel de conocimiento teórico menor sería la dependencia de datos experimentales (Jones et al., 2007; Arendt et al., 2018). Sin embargo, para sistemas complejos (como los que involucran microorganismos) y/o cuya estructura y comportamiento cambia en el tiempo, la complejidad matemática es mayor, lo que limitaría la solución y uso del modelo matemático (Bernard et al., 2006). Como consecuencia de lo anterior, estimar el valor "verdadero" de los parámetros se convierte en un reto aún más complejo.

Particularmente, los modelos de matemáticos de procesos biotecnológicos tienen como objetivo describir la fenomenología, de forma determinística, que exhiben microorganismos o parte de los mismos en cuanto a transporte, reacción química y comportamiento social. Sin embargo, en la realidad estos procesos cuentan con un componente estocástico en su naturaleza debido a factores como variabilidad biológica intrínseca, adaptaciones específicas, perturbaciones ambientales, entre otros (Allen, 2010; Bressloff, 2014). Para que el modelo matemático describa entonces de manera exitosa el comportamiento del sistema biológico, requiere de parámetros que condensen la información tanto estocástica como determinística del sistema en un único valor (Nickel et al., 2017).

Para llegar a conocer el valor estimado de un parámetro es necesario primero saber si este es, de hecho, factible de ser conocido. Esta problemática es abordada por los análisis de ***identificabilidad estructural*** de parámetros. Estos análisis se basan únicamente en información de la estructura del modelo (conformada por los estados, parámetros, entradas y ecuaciones que los relacionan) junto con las salidas capaces de cuantificarse experimentalmente, para determinar el tipo de identificabilidad de los parámetros. Un parámetro pueden ser catalogado como: ***globalmente identificable*** si para una estructura y un conjunto de salidas observadas, el parámetro tiene un único valor. Por otro lado, es ***localmente identificable*** si para una estructura y un conjunto de salidas observadas el parámetro tiene un conjunto finito de soluciones. Finalmente, se tienen los parámetros ***no identificables***, los cuales para una estructura y un conjunto de salidas observadas poseen infinitas soluciones

(DiStefano III, 2015).

Cuando existe por lo menos un parámetro no identificable, se dice que el modelo es no identificable, y por tanto, existe incertidumbre total en al menos un estado descrito por el modelo. Lo anterior implica que el modelo carecerá tanto de capacidad descriptiva como predictiva (Villaverde & Banga, 2017). El análisis de identificabilidad estructural solo es redundante en el caso de que todos los estados del modelo puedan ser cuantificados experimentalmente. Sin embargo y manera general para modelos de procesos biotecnológicos, la cantidad de salidas experimentales es menor que los estados modelados, por lo cual este análisis es requerido. Esto se ha demostrado para los modelos descritos en la Tabla **3-1**, tomada de (Villaverde et al., 2016).

Tabla 3-1: Análisis de identificabilidad de algunos modelos matemáticos de procesos biotecnológicos, tomada de (Villaverde et al., 2016).

Modelo	Estados	Parámetros	Salidas observadas	¿Identificable estructuralmente?
Consumo hepático de pitavastatina	3	7	1	Si
Oscilador de Goodwin	3	1	1	No
Cascada MAPK con retroalimentación mixta	3	14	3	Si
Ruta metabólica NF-κB	15	29	6	No
Ruta metabólica JAK/STAT	10	23	8	No
Reloj circadiano de *A. thaliana*	7	28	2	No
Modelo metabólico de células CHO	34	117	13	No

La estimación de parámetros realizada de manera rigurosa requiere entonces de diversas estrategias que deben aplicarse de manera sistemática para la consecución de parámetros con un significado físico y biológico acorde a la realidad, y por ende, extensible a la filosofía de Ingeniería de Sistemas de Proceso (PSE, por sus siglas en inglés) (Cameron & Hangos, 2001). De esta manera, si el modelo es *identificable* entonces el valor de sus parámetros puede ser conocido, y por tanto, brindar información útil en aplicaciones de PSE. Así mismo, estimar el valor de los parámetros implica un esfuerzo considerable dado que requiere de la solución de un problema de optimización, en consecuencia, si los parámetros no son identificables derivaría en un uso ineficaz de recursos.

3.3. Metodología

La estimación estructural de parámetros consiste en determinar si el valor de un parámetro es susceptible de ser estimado considerando la estructura del modelo (que involucra sus estados, parámetros, entradas y ecuaciones) junto con las salidas observadas experimentalmente. El concepto de identificabilidad estructural surge de la observabilidad del modelo matemático, es decir, si es teóricamente posible inferir los estados no observados del modelo mediante la observación de algunas salidas del mismo. En este caso, la observabilidad se extiende a la capacidad de estimar los parámetros a través de la observación de algunas salidas del modelo (Villaverde, 2019). El concepto de observabilidad de sistemas fue introducido por Kalman *et al.* en 1960 para sistemas lineales invariantes en el tiempo (Kalman, 1960) y posteriormente extendido para sistemas no lineales (Kostyukovskii, 1968; Griffith & Kumar, 1971).

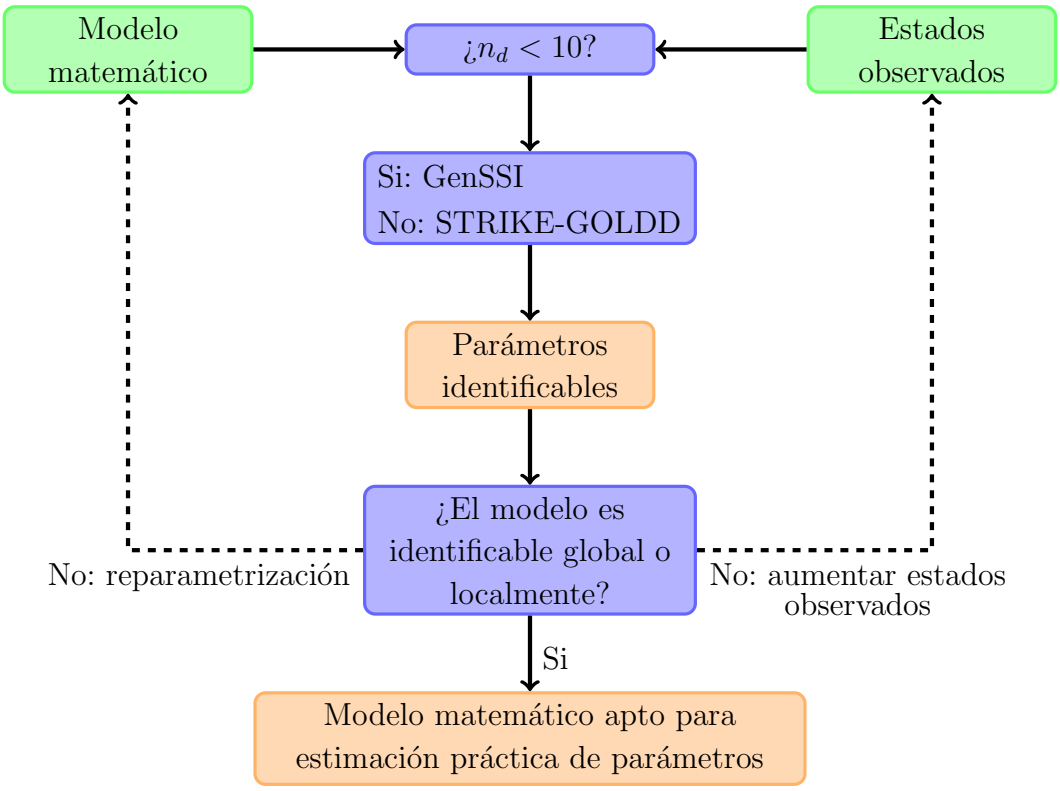


Figura 3-1: Metodología de análisis de identificabilidad estructural.

La Figura **3-1** presenta la metodología utilizada en esta investigación para el análisis de identificabilidad estructural del modelo matemático de fermentación diaúxica para producción de xilitol. En caso de que el modelo no sea identificable estructuralmente, existen dos

caminos posibles: reparametrización o aumento en el número de estados observados.

La reparametrización implica la modificación de la estructura del modelo en favor de combinaciones identificables creadas a partir de parámetros no identificables. Estas combinaciones pueden ser encontradas a través de diversos métodos, como la implementación basada en geometría diferencial programa en el paquete STRIKE-GOLDD para Matlab® (Villaverde et al., 2016; Massonis & Villaverde, 2020), un método de perfilación de subconjuntos de parámetros basado en la matriz de información de Fisher (Eisenberg & Hayashi, 2014) y un método basado en álgebra diferencial llamado "COMBOS" con implementación web (Meshkat et al., 2009, 2012, 2014).

Por otra parte, el aumento del número de estados observados implica la modificación del diseño experimental, al requerir el muestreo de más variables del sistema estudiado. Sin embargo, esta aproximación puede ser imposible o sumamente difícil en algunos casos, por ejemplo, la cuantificación de concentraciones en orgánulos intracelulares. No obstante, el análisis de identificabilidad estructural dependiente de entradas propuesto por Villaverde *et al.*, puede ayudar a determinar la cantidad de experimentos (entradas constantes) o el tipo de experimento (entradas variables) necesario para lograr identificabilidad estructural (Villaverde et al., 2019).

3.3.1. Modelo matemático

El caso de estudio seleccionado corresponde al modelo matemático de bioproducción de xilitol, propuesto por Tochampa *et al.* (Tochampa et al., 2015) mostrado brevemente en las Ecuaciones 3-1 a 3-7. Este modelo cuenta con 5 estados y 11 parámetros, además de considerar fenómenos como inhibición y transporte de metabolitos.

$$\frac{dC_X}{dt} = \mu C_X \tag{3-1}$$

$$\frac{dC_{glu}}{dt} = -\left[q_{glu}^{max} \frac{C_{glu}}{C_{glu} + K_{S,glu}\left(1 + \frac{C_{xil}}{K_{i,xil}}\right)} \right] C_X \tag{3-2}$$

$$\frac{dC_{xil}}{dt} = -\left[q_{xil}^{max} \frac{C_{xil}}{C_{xil} + K_{S,xil}\left(1 + \frac{C_{glu}}{K_{i,glu}}\right)} \right] C_X \tag{3-3}$$

$$\frac{dC_{xit}^{in}}{dt} = \rho_X(r_{f,xit} - r_{u,xit} - r_{t,xit}) - \mu C_{xit}^{in} \tag{3-4}$$

$$\frac{dC_{xit}^{ex}}{dt} = r_{t,xit}C_X \tag{3-5}$$

$$\mu = \mu_{glu}^{max}\frac{C_{glu}}{K_{S,glu} + C_{glu}} + \mu_{xit}^{max}\frac{C_{xit}^{in}}{K_{S,xit} + C_{xit}^{in}}\frac{K_r}{K_r + C_{glu}} \tag{3-6}$$

$$r_{t,xit} = 3.6\text{x}10^6 P_{xit}a_{cell}(C_{xit}^{in} - C_{xit}^{ex}) \tag{3-7}$$

3.3.2. Metodologías de identificabilidad estructural de parámetros

Diferentes aproximaciones han sido planteadas para determinar la observabilidad de parámetros en modelos matemáticos. Entre ellas se encuentran los métodos basados en series de potencias, álgebra diferencial y geometría diferencial.

- **Series de potencias**

En el método propuesto por Pohjanpalo, se considera el siguiente sistema de ecuaciones diferenciales (Pohjanpalo, 1978):

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t, \boldsymbol{\theta}), \qquad \mathbf{x}(t) \in R^n,\ t \in [0, T] \\ \mathbf{y}(t) &= \mathbf{g}(\mathbf{x}(t), \boldsymbol{\theta}) \end{aligned} \tag{3-8}$$

en donde $\mathbf{x}$ corresponde al vector de estados, $\boldsymbol{\theta}$ al vector de parámetros del modelo, $\mathbf{u}$ vector de entradas, $\mathbf{y}$ vector de estados observados, $\mathbf{f}$ vector de ecuaciones diferenciales de los estados y $\mathbf{g}$ vector de funciones de mapeo de estados medidos. Asumiendo que las funciones $\mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t, \boldsymbol{\theta})$ y $\mathbf{g}(\mathbf{x}(t), \boldsymbol{\theta})$ son continuas en todo su rango, y por ende, infinitamente diferenciables respecto a $\mathbf{x}$ y t, se puede afirmar que dichas funciones cumplen la condición de Lipschitz y tendrán una única solución. Las derivadas entonces pueden notarse como:

$$a_k \stackrel{\triangle}{=} y^{(k)} \tag{3-9}$$

en donde k indica la k-esima derivada respecto al tiempo.

La identificabilidad estructural del modelo se puede establecer si el siguiente conjunto de ecuaciones tiene única solución para θ:

$$\mathbf{g}^{(k)}(\mathbf{x}(0), \theta) = a_k(0), \qquad k = 0, ..., \infty \tag{3-10}$$

Lo anterior se debe a que al cumplirse la condición de Lipschitz, la función tendrá una única trayectoria que dependerá de los parámetros y las condiciones iniciales del sistema. Esta información está contenida en el *germen* (conjunto infinito de sus derivadas en tiempo cero).

Sin embargo, si las derivadas de la función no son definidas en tiempo cero, el modelo no puede considerarse como no identificable. Esto se debe a que observaciones en un tiempo mayor a cero pueden ser informativas para establecer la identificabilidad estructural de los parámetros.

- **Álgebra diferencial**

En el método propuesto por Saccomani et al. (2001), se consideran el sistema definido por la Ecuación 3-8 y la función $\Phi(p, u)$ (Ecuación 3-11) que corresponde al mapa de entradas y salidas del sistema. El sistema es *a priori globalmente identificable*, sí y solo sí, por lo menos para un conjunto de valores $\boldsymbol{\theta}^* \in \Theta$ se cumple que:

$$\Phi(\boldsymbol{\theta}^*, \mathbf{u}) = \Phi(\boldsymbol{\theta}, \mathbf{u}) \tag{3-11}$$

y $\boldsymbol{\theta}^* = \boldsymbol{\theta}$ tiene una única solución. Si existe un conjunto finito de soluciones para $\boldsymbol{\theta}$ el sistema es *a priori localmente identificable* y es *no identificable* si existe un conjunto infinito de soluciones.

El modelo descrito por la Ecuación 3-8 puede expresarse como un conjunto de $n + r$ polinomios diferenciales:

$$\dot{\mathbf{x}}(t) - \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t, \boldsymbol{\theta}) \tag{3-12}$$
$$\mathbf{y}(t) - \mathbf{g}(\mathbf{x}(t), \boldsymbol{\theta}) \tag{3-13}$$

estos polinomios son los generadores de un *ideal diferencial* I en un anillo diferencial. El conjunto característico de un ideal I es un conjunto finito de $n + r$ ecuaciones diferenciales no lineales que describe el mismo conjunto solución del sistema original. Para obtener dicho conjunto característico de polinomios diferenciales se requiere de un *anillo diferencial* $R(\boldsymbol{\theta})[\mathbf{x}, \mathbf{y}, \mathbf{u}]$, en donde $R(\boldsymbol{\theta})$ es un campo de funciones racionales de los parámetros y tiene como variables los estados, las salidas observadas y las entradas. Este conjunto de polinomios racionales se obtiene a través del algoritmo de pseudo-división de Ritt (Meshkat et al., 2012).

Se debe realizar una clasificación de las variables del anillo diferencial, normalmente se toma lo siguiente:

$$\begin{aligned} \mathbf{u}_1 <& \dot{\mathbf{u}}_1 < \ldots < \mathbf{u}_2 < \dot{\mathbf{u}}_2 < \ldots < \mathbf{y}_1 < \dot{\mathbf{y}}_1 \ldots \\ & \ldots \mathbf{y}_2 < \dot{\mathbf{y}}_2 < \ldots < \mathbf{x}_1 < \dot{\mathbf{x}}_1 < \ldots < \mathbf{x}_2 < \dot{\mathbf{x}}_2 < \ldots \end{aligned} \tag{3-14}$$

entonces, el conjunto característico de polinomios diferenciales corresponde a las relaciones entrada-salida, y tiene la siguiente forma:

$$\begin{aligned} &A_1(\mathbf{u},\mathbf{y})\dots A_r(\mathbf{u},\mathbf{y}) \\ &A_{r+1}(\mathbf{u},\mathbf{y},x_1) \\ &A_{r+2}(\mathbf{u},\mathbf{y},x_1,x_2) \\ &\vdots \\ &A_{r+n}(\mathbf{u},\mathbf{y},x_1,\dots,x_n) \end{aligned} \tag{3-15}$$

en donde

$$A_l(\mathbf{u},\mathbf{y}) = \sum_{i=1}^{r+n} c_i(\theta) M_i \tag{3-16}$$

c_i corresponde a una función racional de los parámetros θ y M_i a "monomiales" o productos de potencias de los estados, salidas observadas o entradas. Para las primeras r ecuaciones de polinomios diferenciales se tiene:

$$A_1(\mathbf{u},\mathbf{y}) = 0\,, A_2(\mathbf{u},\mathbf{y}) = 0\,, \dots\,, A_r(\mathbf{u},\mathbf{y}) = 0 \tag{3-17}$$

Las ecuaciones anteriores son entonces el anillo diferencial $R(\theta)[u,y]$. Al resolver el sistema de ecuaciones formado por el conjunto de Ecuaciones 3-17 con $c_i(\theta)$ en términos de los parámetros, la cantidad de soluciones determinará entonces el tipo de identificabilidad de los parámetros del modelo matemático. El análisis de identificabilidad estructural basado en álgebra diferencial ha sido implementado en el software DAISY (Saccomani & D'angiò, 2009).

- **Geometría diferencial**

El análisis de identificabilidad estructural basado en geometría diferencial (Villaverde, 2019), requiere de la noción de estados distinguibles. Dos estados son *indistinguibles* si $y_{x_1}(t) = y_{x_2}(t)$, en donde $y(t)$ es la evolución en el tiempo de la salida de un modelo que parte de un estado x_0 en un tiempo t_0. También, requiere de la noción de *observabilidad*, la cual describe la posibilidad de determinar un estado basado en mediciones presentes y futuras.

Para el caso de sistemas no lineales como el descrito por la Ecuación 3-8, se requiere el uso de derivadas de Lie, las cuales se definen como:

$$L_f \mathbf{g}(x) = \frac{\partial \mathbf{g}(\mathbf{x})}{\partial \mathbf{x}} \mathbf{f}(\mathbf{x}) \tag{3-18}$$

y las derivadas de Lie de orden superior pueden ser calculadas de manera recursiva como:

$$\begin{aligned} L_f^2\mathbf{g}(\mathbf{x}) &= \frac{\partial L_f\mathbf{g}(\mathbf{x})}{\partial \mathbf{x}}\mathbf{f}(\mathbf{x}) \\ &\vdots \\ L_f^i\mathbf{g}(\mathbf{x}) &= \frac{\partial L_f^{i-1}\mathbf{g}(\mathbf{x})}{\partial \mathbf{x}}\mathbf{f}(\mathbf{x}) \end{aligned} \tag{3-19}$$

Tomando en cuenta lo anterior, se puede definir la matriz de observabilidad de un sistema no lineal $\mathscr{O}^{NL}$ como:

$$\mathscr{O}^{NL}(\mathbf{x}) = \begin{pmatrix} \frac{\partial}{\partial \mathbf{x}}\mathbf{g}(\mathbf{x}) \\ \frac{\partial}{\partial \mathbf{x}}L_f\mathbf{g}(\mathbf{x}) \\ \frac{\partial}{\partial \mathbf{x}}L_f^2\mathbf{g}(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x}L_f^{n-1}\mathbf{g}(\mathbf{x}) \end{pmatrix} \tag{3-20}$$

Si el modelo no lineal satisface el rango$\big(\mathscr{O}^{NL}(\mathbf{x}_0)\big) = n$, donde n corresponde a la cantidad total de estados, entonces el modelo es localmente observable alrededor de $\mathbf{x}_0$.

La identificabilidad estructural del modelo no lineal puede entenderse como un caso particular de observabilidad, en donde los parámetros son considerados como estados sin dinámica, entonces el vector de estados aumentado es:

$$\tilde{\mathbf{x}} = \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\theta} \end{bmatrix} \tag{3-21}$$

Se redefine la matriz de observabilidad $\mathscr{O}^{NL}$ para el vector de estados aumentado:

$$\mathscr{O}^{NL}(\tilde{\mathbf{x}}) = \begin{pmatrix} \frac{\partial}{\partial \tilde{\mathbf{x}}}\mathbf{g}(\tilde{\mathbf{x}}) \\ \frac{\partial}{\partial \tilde{\mathbf{x}}}L_f\mathbf{g}(\tilde{\mathbf{x}}) \\ \frac{\partial}{\partial \tilde{\mathbf{x}}}L_f^2\mathbf{g}(\tilde{\mathbf{x}}) \\ \vdots \\ \frac{\partial}{\partial \tilde{\mathbf{x}}}L_f^{n+q-1}\mathbf{g}(\tilde{\mathbf{x}}) \end{pmatrix} \tag{3-22}$$

en donde q indica la cantidad total de parámetros que posee el modelo. Si el rango$\big(\mathscr{O}^{NL}(\tilde{x})\big) = n+q$, entonces el modelo es localmente observable e identificable en una vecindad $\mathscr{N}(\tilde{x}_0)$ de $\tilde{x}_0$. Dado que cada columna de $\mathscr{O}^{NL}$ corresponde a la derivada parcial respecto a un estado o parámetro, es posible determinar los parámetros no identificables removiendo la columna correspondiente y recalculando el rango de $\mathscr{O}^{NL}$. Si al eliminarse la i-esima columna el rango de $\mathscr{O}^{NL}$ no cambia, entonces el i-esimo parámetro es estructuralmente no identificable.

3.3.3. Herramientas computacionales

Actualmente existen dos *toolboxes* creados en el software Matlab® R2018b destinados a la identificabilidad estructural de parámetros: GenSSI y STRIKE-GOLDD (Chiş et al., 2011a; Villaverde et al., 2016). El primero de estos se basa en el uso de generación de series acoplado con tablas de identificabilidad y el segundo en geometría diferencial. Los códigos de estos *toolboxes* son de uso libre, sin embargo, al ser escritos en Matlab se requiere licencia de este último.

- **GenSSI**

El *toolbox* GenSSI (Generating Series for testing Structural Identifiability) fue concebido por Chiş y colaborares en 2011 (Chiş et al., 2011a). Esta aplicación se basa en la generación de series acoplada con tablas de identificabilidad. Esta herramienta requiere de ingresar un archivo de Matlab® R2018b especificando los estados, parámetros y entradas del modelo de manera simbólica junto con los vectores de parámetros, estados, entradas, ecuaciones diferenciales, condiciones iniciales y salidas observadas, según un *template* especificado.

Las series se construyen a través de la solución de un sistema no lineal de ecuaciones, en términos de los parámetros del modelo definido en las Ecuaciones 3-1 a 3-7, mediante la computación de derivadas de Lie sucesivas. Posteriormente, se obtienen las tablas de identificabilidad, que corresponden a los elementos no nulos del jacobiano de los coeficientes de las series generadas. El jacobiano cuenta con un número de columnas igual a la cantidad de parámetros y una cantidad de filas correspondiente a coeficientes no nulos de las series, en principio infinitas (Chiş et al., 2011b).

El método empleado por GenSSI es muy poderoso, sin embargo, no es eficiente si el análisis requiere derivadas de Lie de séptimo orden o mayor, dado que el tiempo de cómputo crece exponencialmente con cada orden de estas derivadas. La Figura **3-2** muestra el resultado obtenido para el modelo ruta metabólica de glucólisis, el cual cuenta con 5 estados, 5 parámetros, 4 entradas y 5 salidas observadas.

La Figura **3-2**(a) presenta la tabla de identificabilidad obtenida como resultado del cálculo de las series. La Figura **3-2**(b) muestra la tabla de identificabilidad reducida, en la cual solo se consideran las primeras filas linealmente independientes con las que puede establecerse la identificabilidad del modelo. La Figura **3-2**(c) corresponde a la solución simbólica de los parámetros. La Figura **3-2**(d) corresponde a la conclusión de identificabilidad estructural del modelo analizado.

El eje vertical de las tablas de identificabilidad indica las series generadas y en el eje hori-

zontal se muestran los parámetros del modelo. Los cuadros negros indican interacciones no nulas entre los diferentes parámetros para una misma serie. Si existen dichas interacciones es posible entonces resolver el sistema de ecuaciones generado y encontrar las soluciones para los parámetros y el criterio de identificabilidad. En el caso que existan columnas completamente blancas en la tabla de identificabilidad, significaría que el modelo no es sensible a ese parámetro en particular, y por tanto, el mismo es no identificable. Tomando en cuenta la Figura **3-2**(d), el modelo de ruta metabólica de glucólisis con 5 salidas observadas es globalmente identificable.

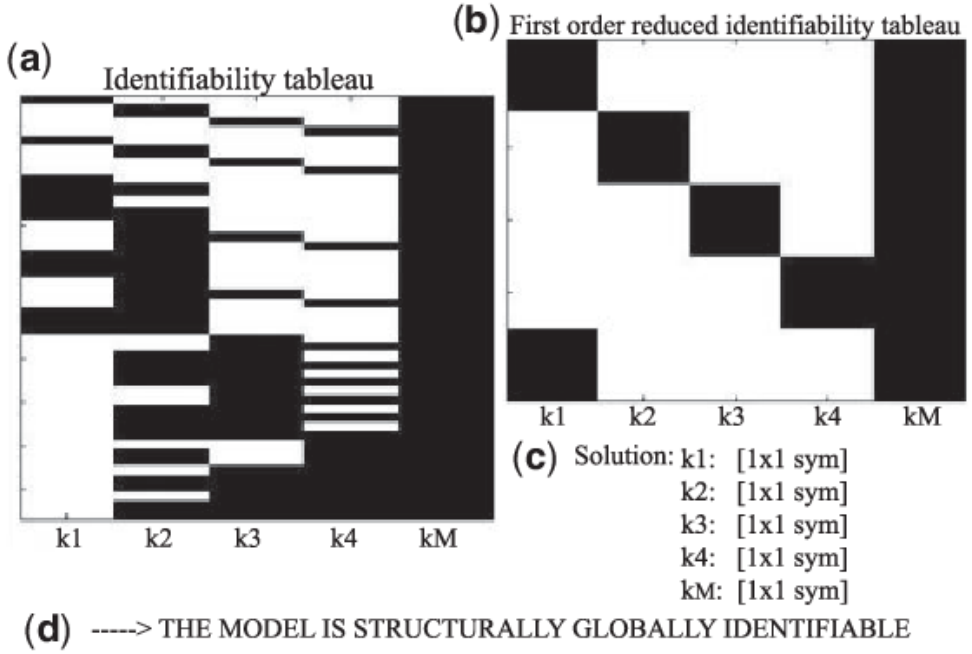


Figura 3-2: Ejemplo de resultado obtenido con GenSSI, tomado de Chiş et al. (2011a). (a) Tabla de identificabilidad, (b) Tabla de identificabilidad reducida (mínimo número de filas necesarias para establecer un criterio de identificabilidad), (c) Resultado obtenido para cada parámetro, (d) Criterio de identificabilidad del modelo matemático analizado.

- **STRIKE-GOLDD**

El *toolbox* STRIKE-GOLDD (STRuctural Identifibalitiy taKen as Extended-Generalized Observability using Lie Derivatives and Decompositiion) concebido e implementado por Villaverde et al. (2016) se basa en la aproximación de geometría diferencial y observabilidad del modelo matemático. El programa se basa inicialmente en la siguiente relación:

$$n_d = \left[\frac{n+q}{m} - 1\right] \tag{3-23}$$

en donde n_d indica el mínimo número de derivadas de Lie que deben ser calculadas para alcanzar rango completo y n, q y m que corresponden a los números de estados, parámetros y salidas observadas, respectivamente. Una vez esta cantidad de derivadas de Lie ha sido calculada, la adición de cada nueva derivada de Lie es seguida del cálculo del rango de la matriz de identificabilidad-observabilidad.

Sí el número de derivadas de Lie es muy alto ($n_d \geq 10$), el programa tiene la opción de descomponer el modelo en submodelos para facilitar el análisis. Los submodelos son encontrados a través de optimización combinatorial en donde se minimiza n_d:

$$\min_{\mathrm{s}} n_d(\mathrm{s}) \tag{3-24}$$

en donde $s = \{s_1, s_2, \ldots, s_n\}$ es un vector binario de tamaño n, en donde $s_j = 1$ denota inclusión y $s_j = 0$ exclusión del estado correspondiente. Dadas las ventajas de esta implementación, se recomienda su uso para modelos matemáticos con un elevado número parámetros.

El paquete STRIKE-GOLDD requiere de un archivo de entrada con las mismas definiciones que el usado para el paquete GenSSI, según el *template* aportado por sus creadores. Así mismo, la salida de este paquete se compone de varios archivos entre los que se encuentra `id_results_MODELNAME_DATE.mat` que contiene las variables p_id (parámetros identificables) y p_un (parámetros no identificables), entre otras. Por otra parte, está el archivo `obs_ident_matrix_MODEL_NUMBER_OF_Lie_deriv.mat` que contiene las matrices de identificabilidad calculadas en el proceso. Finalmente, se tiene el archivo `decomp_MODEL_DATE_MAXSTATES_MAXLIETIME.mat` si fue necesario realizar descomposición del modelo original.

3.4. Resultados

A continuación se muestran los resultados del análisis de identificabilidad estructural del modelo matemático de fermentación diaúxica de glucosa y xilosa para producción de xilitol.

3.4.1. Identificabilidad estructural del modelo de fermentación diaúxica de glucosa y xilosa

El modelo matemático descrito por las Ecuaciones 3-1 a 3-7, cuenta con 5 estados, 11 parámetros y 4 salidas observadas. Por tanto, se tiene

$$n_d = \left[\frac{n+q}{m} - 1\right] = \left[\frac{5+11}{4} - 1\right] = 3$$

Dado que el número mínimo de derivadas de Lie necesarias para establecer la identificabilidad estructural del modelo es 3, es factible obtener un resultado a través del toolbox GenSSI, pues no se requiere la descomposición de este modelo. La información experimental para este modelo fue tomada de la investigación de Sirisansaneeyakul *et al.*, la cual cuenta con 21 experimentos en modalidad batch repetidos con condiciones iniciales diferentes (Sirisansaneeyakul et al., 2013). Se realizó un análisis de identificabilidad estructural en Matlab® 2018b para cada una de estas condiciones, debido a que influyen en la identificabilidad estructural de los parámetros.

La Figura **3-3** muestra el resultado del análisis de identificabilidad para condiciones iniciales de glucosa y xilosa diferentes de cero, específicamente 10.96 g/L de glucosa y 5.15 g/L de xilosa (experimento 1). Al observar la tabla de identificabilidad se evidencia que no existen columnas blancas, lo que indica que el modelo es sensible a todos los parámetros, y por tanto, su identificabilidad puede ser local o global.

En los experimentos restantes se varió la concentración inicial de xilosa, pero no se alimentó glucosa. Para efectos de análisis de identificabilidad, estos experimentos proporcionan la misma información y solo se mostrará el resultado obtenido con el experimento 2 en donde la concentración inicial de xilosa fue 26.62 g/L. La Figura **3-4** muestra columnas blancas para varios parámetros, esto indica que el modelo no es sensible a ellos, y por tanto, son no identificables. Los resultados explícitos del análisis de identificabilidad estructural para ambos casos se muestran en la Tabla **3-2**.

Tabla 3-2: Clasificación de los parámetros del modelo de fermentación diaúxica para producción de xilitol.

Parámetro	Condición inicial	
	Glucosa y Xilosa	Xilosa
μ_{glu}^{max}	Localmente identificable	No identificable
μ_{xit}^{max}	Localmente identificable	Globalmente identificable
q_{glu}^{max}	Localmente identificable	No identificable
q_{xil}^{max}	Localmente identificable	Globalmente identificable
Ks_{xil}	Localmente identificable	Globalmente identificable
Ks_{glu}	Localmente identificable	No identificable
Ks_{xit}	Localmente identificable	Globalmente identificable
Ki_{xil}	Localmente identificable	No identificable
Ki_{glu}	Localmente identificable	No identificable
K_r	Localmente identificable	No identificable
P_{xit}	Localmente identificable	Globalmente identificable
Modelo	Identificable	No identificable

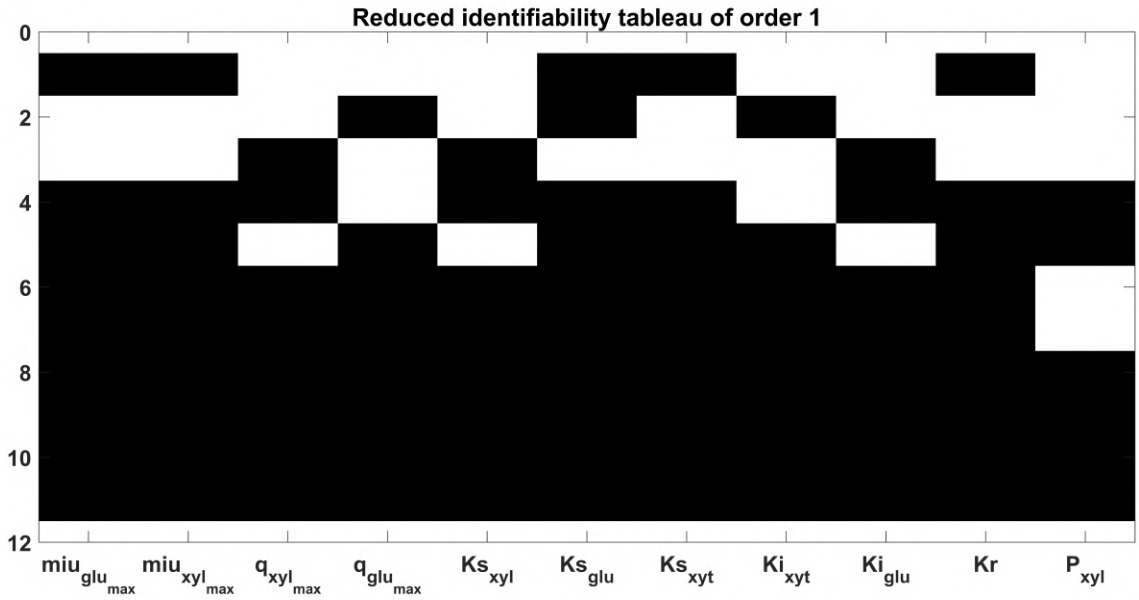


Figura 3-3: Tabla de identificabilidad para el modelo de fermentación diaúxica para producción de xilitol con concentraciones iniciales de glucosa (10.96 g/L) y xilosa (5.15 g/L), obtenido con GenSSI toolbox.

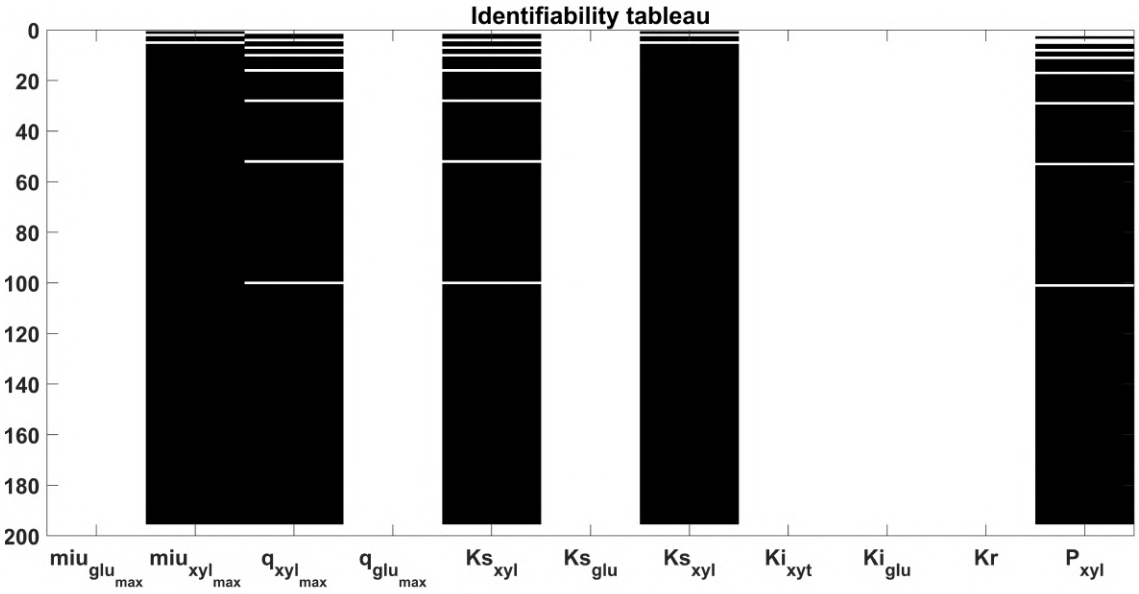


Figura 3-4: Tabla de identificabilidad para el modelo de fermentación diaúxica para producción de xilitol con concentración inicial variable de xilosa y nula de glucosa.

Es de resaltar que si existe una condición inicial con concentración de biomasa nula, todos los parámetros se hacen no identificables. Esto es lógico debido a que sin biomasa no habría reacciones bioquímicas (consumo de sustratos y producción de metabolitos), y por ende, no

hay sensibilidad del modelo respecto a ningún parámetro.

Los resultados mostrados en la Tabla **3-2**, muestran que bajo condiciones iniciales no nulas de ambos sustratos todos los parámetros son localmente identificables. Esto implica que el modelo es sensible a todos los parámetros y que si se cambiaran las condiciones iniciales el valor numérico de los mismos podría cambiar, lo cual estaría relacionado a posibles puntos de bifurcación. No obstante, se encuentra que bajo concentración inicial nula de glucosa, los parámetros asociados a este sustrato se vuelven no identificables, en contraste, los parámetros asociados a xilosa y xilitol ($\mu_{max_{xit}}$, $q_{max_{xit}}$, $K_{s_{xil}}$, $K_{s_{xit}}$ y P_{xit}) se vuelven globalmente identificables. Lo anterior expone una secuencia de estimación de los valores numéricos de los parámetros para este modelo matemático, pues indica los conjuntos de datos a utilizar y los parámetros que pueden obtenerse de cada uno de ellos. De esta manera, se establece un puente entre los componentes y estructura del modelo con la información experimental disponible.

3.5. Conclusiones

La estimación de estructural de parámetros es un análisis fundamental que debe realizarse a todo modelo cuyos parámetros deben ser estimados a partir de datos experimentales, con el fin de garantizar validez en la interpretación del modelo matemático. Debe ser resaltado que sí experimentalmente todos los estados del modelo no pueden ser cuantificados, el análisis de identificabilidad estructural siempre debe ser realizado. Así mismo, este análisis puede ser una herramienta que contribuya tanto a la elaboración de la estructura de un modelo, mediante re-estructuración para eliminar parámetros no identificables, como al diseño experimental requerido (estados a observar). En el caso del modelo de fermentación diaúxica de glucosa y xilosa, se encontró que sí existen concentraciones iniciales diferentes de cero para glucosa y xilosa, los 11 parámetros del modelo son localmente identificables, por tanto, el modelo es identificable desde un punto de vista estructural. Adicionalmente y de manera general, se aconseja seguir la metodología de la Figura **3-1** para el análisis estructural de cualquier modelo matemático de procesos biotecnológicos.

Tabla 3-3: Nomenclatura matemática.

Símbolo	Definición	Símbolo	Definición
$\mathbf{x}(t)$	Vector de estados	$\mathbf{u}(t)$	Vector de entradas
t	Tiempo	$\boldsymbol{\theta}$	Vector de parámetros
$\dot{\mathbf{x}}(\mathbf{t})$	Vector de derivadas de los estados	$\mathbf{f}$	Vector de funciones diferenciales de los estados
$\mathbf{y}(t)$	Vector de estados observados	$\mathbf{g}$	Vector de funciones de mapeo estados observados
a_k	k-esima derivada respecto al tiempo	$g^{(k)}$	k-esima derivada de g
Φ	Función de mapeo de entradas y estados observados	A	Polinomio diferencial
c_i	Función racional de θ	M_i	Monomial de potencias de estados, estados observados o entradas
$L_f g(x)$	Derivada de Lie	$\mathscr{O}^{NL}(x)$	Matriz de observabilidad de estados
$\tilde{\mathbf{x}}$	Vector de estados aumentado	$\mathscr{O}^{NL}(\tilde{\mathbf{x}})$	Matriz de observabilidad de estados aumentada

Bibliografía

Allen, L. J. (2010). An introduction to stochastic processes with applications to biology. CRC Press.

Arendt, K., Jradi, M., Shaker, H. R., & Veje, C. (2018). Comparative analysis of white-, gray- and black-box models for thermal simulation of indoor environment: Teaching building case study. In Proceedings of the 2018 Building Performance Modeling Conference and SimBuild co-organized by ASHRAE and IBPSA-USA, Chicago, IL, USA (pp. 26–28).

Bernard, O., Chachuat, B., Hélias, A., & Rodriguez, J. (2006). Can we assess the model complexity for a bioprocess: theory and example of the anaerobic digestion process. Water science and technology, 53(1), 85–92.

Biegler, L. T. (2010). Nonlinear programming: concepts, algorithms, and applications to chemical processes, volume 10. Siam.

Bressloff, P. C. (2014). Stochastic processes in cell biology, volume 41. Springer.

Cameron, I. T. & Hangos, K., Eds. (2001). Process Modelling and Model Analysis, volume 4. Academic press.

Chiş, O., Banga, J. R., & Balsa-Canto, E. (2011a). Genssi: a software toolbox for structural identifiability analysis of biological models. Bioinformatics, 27(18), 2610–2611.

Chiş, O., Banga, J. R., & Balsa-Canto, E. (2011b). Methods for checking structural identifiability of nonlinear biosystems: A critical comparison. IFAC Proceedings Volumes, 44(1), 10585–10590.

DiStefano III, J. (2015). Dynamic systems biology modeling and simulation. Academic Press.

Eisenberg, M. C. & Hayashi, M. A. (2014). Determining identifiable parameter combinations using subset profiling. Mathematical biosciences, 256, 116–126.

Englezos, P. & Kalogerakis, N. (2000). Applied parameter estimation for chemical engineers. CRC Press.

Griffith, E. W. & Kumar, K. (1971). On the observability of nonlinear systems: I. Journal of Mathematical Analysis and Applications, 35(1), 135–147.

Jones, D., Watton, J., & Brown, K. (2007). Comparison of black-, white-, and grey-box models to predict ultimate tensile strength of high-strength hot rolled coils at the port talbot hot strip mill. Proceedings of the Institution of Mechanical Engineers, Part L: Journal of Materials: Design and Applications, 221(1), 1–9.

Kalman, R. E. (1960). On the general theory of control systems. In Proceedings First International Conference on Automatic Control, Moscow, USSR.

Kostyukovskii, Y. M. (1968). Simple conditions of observability of nonlinear controlled systems. Avtomat. Telemekh, 10, 32–41.

Massonis, G. & Villaverde, A. F. (2020). Finding and breaking lie symmetries: Implications for structural identifiability and observability in biological modelling. Symmetry, 12(3), 469.

Meshkat, N., Anderson, C., & DiStefano III, J. J. (2012). Alternative to ritt's pseudodivision for finding the input-output equations in algebraic structural identifiability analysis. Mathematical Biosciences, 239(1), 117–123.

Meshkat, N., Eisenberg, M., & DiStefano III, J. J. (2009). An algorithm for finding globally identifiable parameter combinations of nonlinear ode models using gröbner bases. Mathematical biosciences, 222(2), 61–72.

Meshkat, N., Kuo, C. E.-z., & DiStefano III, J. (2014). On finding and using identifiable parameter combinations in nonlinear dynamic systems biology models and combos: a novel web implementation. PLoS One, 9(10).

Nickel, D. B., Cruz-Bournazou, M. N., Wilms, T., Neubauer, P., & Knepper, A. (2017). Online bioprocess data generation, analysis, and optimization for parallel fed-batch fermentations in milliliter scale. Engineering in Life Sciences, 17(11), 1195–1201.

Pohjanpalo, H. (1978). System identifiability based on the power series expansion of the solution. Mathematical biosciences, 41(1-2), 21–33.

Saccomani, M. P., Audoly, S., Bellu, G., & D'Angio, L. (2001). A new differential algebra algorithm to test identifiability of nonlinear systems with given initial conditions. In Proceedings of the 40th IEEE Conference on Decision and Control, volume 4 (pp. 3108–3113).: IEEE.

Saccomani, M. P. & D'angiò, L. (2009). Examples of testing global identifiability with the daisy software. IFAC Proceedings Volumes, 42(10), 48–53.

Sirisansaneeyakul, S., Wannawilai, S., & Chisti, Y. (2013). Repeated fed-batch production of xylitol by candida magnoliae tistr 5663. Journal of Chemical Technology & Biotechnology, 88(6), 1121–1129.

Tochampa, W., Sirisansaneeyakul, S., Vanichsriratana, W., Srinophakun, P., Bakker, H. H., Wannawilai, S., & Chisti, Y. (2015). Optimal control of feeding in fed-batch production of xylitol. Industrial & Engineering Chemistry Research, 54(7), 1992–2000.

Villaverde, A. F. (2019). Observability and structural identifiability of nonlinear biological systems. Complexity, 2019.

Villaverde, A. F. & Banga, J. R. (2017). Dynamical compensation and structural identifiability of biological models: Analysis, implications, and reconciliation. PLoS computational biology, 13(11), e1005878.

Villaverde, A. F., Barreiro, A., & Papachristodoulou, A. (2016). Structural identifiability of dynamic systems biology models. PLoS computational biology, 12(10), e1005153.

Villaverde, A. F., Evans, N. D., Chappell, M. J., & Banga, J. R. (2019). Input-dependent structural identifiability of nonlinear systems. IEEE Control Systems Letters, 3(2), 272–277.

Capítulo 4

Metodología para sintonización de algoritmos de optimización global

4.1. Resumen

Los procesos biotecnológicos pueden ofrecer alternativas más eficientes y ecológicamente amigables a los procesos industriales convencionales, que pueden ayudar a mitigar a influencia antrópica adversa en el planeta. Para alcanzar este propósito, la ingeniería de sistemas de proceso (PSE por sus siglas en inglés) reúne un conjunto de herramientas basadas en modelos y métodos que pueden contribuir significativamente a alcanzar sostenibilidad en recursos naturales, por medio de la optimización del diseño y operación de procesos biotecnológicos. Por lo tanto, modelos matemáticos que representen apropiadamente el comportamiento de los sistemas biológicos son necesarios. Generalmente, estos modelos requieren de estimación de sus parámetros usando datos experimentales a través de la solución de un problema de optimización, normalmente no convexo. Las metaheurísticas pueden ser una solución conveniente al problema de optimización, sin embargo, estos optimizadores presentan parámetros internos que necesitan ser sintonizados.

En este capítulo se presenta un análisis de los efectos de la sintonización de parámetros de los optimizadores globales incluidos en el *global optimization toolbox* de Matlab® R2018 y el tipo de normalización de función objetivo para la estimación de parámetros de bioproducción de xilitol. Una herramienta híbrida para la sintonización de algoritmos de optimización fue desarrollada por interconexión entre Matlab® R2018 y R® v4.0.3. La comparación del desempeño entre optimizadores se realizó mediante la prueba estadística no paramétrica de Friedman y la prueba de Dunn para comparaciones múltiples con corrección de Bonferroni. Se encontró que el proceso de sintonización tiene un efecto significativo en el desempeño del optimizador, mejorando la precisión y reproducibilidad del valor de la función objetivo debido a una correcta relación exploración-explotación entre los parámetros del optimizador. Específicamente, la combinación de optimizador Enjambre de Partículas y factor de

normalización por media del estado experimental mostró los mejores resultados entre las combinaciones de optimizadores y factores de normalización de función objetivo considerados en este estudio. La contribución "Assessment of Metaheuristic-Optimization Algorithms Tuning for Parameter Estimation of Xylitol Fermentation Kinetics" en proceso de sometimiento a la revista Biotechnology Journal presenta los resultados de este capítulo.

4.2. Introducción

El continuo aumento de la población a nivel mundial ha llevado a un mayor consumo de recursos, cuya disponibilidad a mediano y largo plazo se ha visto amenazada (Desing et al., 2020). Bajo este escenario, el concepto de sostenibilidad entendida de forma simple como la capacidad de satisfacer las necesidades de la humanidad tanto a corto como largo plazo, ha sido ampliamente estudiado y aplicado a diferentes sectores como el energético, ambiental, económico entre otros (Ben-Eli, 2018; Wang et al., 2018b).

La biotecnología puede ofrecer soluciones para generar procesos industriales más sostenibles, a través de implementación de tecnologías más eficientes basadas en sistemas biológicos (o parte de estos) que puedan reemplazar o complementar la tecnología ya existente (Montana-Hoyos & Fiorentino, 2016). Sin embargo, es evidente desde la literatura que muchos de proceso biotecnológicos tienen baja rentabilidad o son tecno-económicamente inviables con la actual tecnología. Por lo tanto, la optimización de procesos es muy relevante para validar las oportunidades de los bioprocesos para sustituir los procesos convencionales. Desde el punto de vista de ingeniería de sistemas de proceso (PSE por sus siglas en inglés), comportamiento de los sistemas biológicos debe ser expresado en términos útiles para diseño, optimización y control de procesos (Cameron & Hangos, 2001; Stephanopoulos & Reklaitis, 2011). No obstante, los modelos matemáticos de procesos biotecnológicos usualmente pertenecen a la categoría de "caja gris", lo que implica la presencia de parámetros que deben ser estimados a partir de datos experimentales (Koutinas et al., 2012).

La estimación de parámetros es un proceso computacionalmente intensivo que requiere de un algoritmo de optimización que encuentre los valores de los parámetros que mejor aproximen las predicciones del modelo a los datos experimentales utilizados (Villaverde et al., 2019). Los algoritmos de optimización pueden dividirse en determinísticos y estocásticos. Los algoritmos estocásticos se diferencian de los determinísticos en el hecho de que utilizan decisiones aleatorias durante el proceso de búsqueda del valor óptimo. La aleatoriedad introducida por este tipo de algoritmos presenta como ventaja la generación de movimientos aleatorios que llevan a la exploración de zonas no analizadas del espacio de búsqueda, lo cual puede evitar el estancamiento del algoritmo en un mínimo local (Spall, 2012).

De manera general, se tiene que el espacio de búsqueda definido para los parámetros debe ser

finito y el optimizador debe hacer seguimiento del mejor resultado obtenido en cada iteración. Lo anterior implica que a medida que aumenta el número de pasos aleatorios, el algoritmo se aproxima a la solución óptima siendo generalmente insensible respecto a la forma de la función objetivo (Zhigljavsky & Zilinskas, 2007). En términos probabilísticos, cuando el número de pasos aleatorios tiende a infinito, la probabilidad de encontrar el mínimo global se aproxima al 100 %. Sin embargo, la solución a un problema de optimización debe obtenerse en un número de pasos (o tiempo) finito, además, es deseable que sea obtenida en el menor tiempo (o número de pasos) posible. Lecchini-Visintini y colaboradores han demostrado que pueden generarse algoritmos de optimización estocásticos con garantía teórica de convergencia hacia el mínimo global en un tiempo finito, en este caso templado simulado (Lecchini-Visintini et al., 2010). Adicionalmente, Sun y colaboradores determinaron esta propiedad en variantes del algoritmo enjambre de partículas (Sun et al., 2012).

No obstante, la aleatoriedad en el recorrido a través del espacio de búsqueda puede derivar en que el algoritmo no llegue a la solución óptima. Esto implica dos problemas, por una parte no se podría saber *a priori* si una solución corresponde al verdadero mínimo global, por otra parte, sí el resultado es reproducible. Los algoritmos de optimización (no solo estocásticos) pueden depender de parámetros que controlen tanto el avance como la convergencia hacia una solución óptima. Ejemplos de estos parámetros incluyen los operadores de fuerzas de evolutivas en el algoritmo genético (Kramer, 2017), temperatura inicial y esquema de enfriado en templado simulado (Aguiar e Oliveira et al., 2012), coeficientes sociales y cognitivos, tamaño del enjambre o peso de inercia en métodos de inteligencia de enjambre (Couceiro & Ghamis, 2016), entre otros.

Una forma de mejorar la reproducibilidad del resultado y la calidad de la solución obtenida es la sintonización del algoritmo de optimización. Esto corresponde a un proceso de selección de los parámetros del optimizador basado en el valor alcanzado de la función objetivo. En general, los métodos de sintonización pueden clasificarse en tres categorías: métodos simples de generación-evaluación, métodos de alto nivel de generación-evaluación y métodos iterativos de generación-evaluación. Los métodos simples se basa en la generación y evaluación del conjunto total de posibles configuraciones, los métodos de alto nivel se basan en el uso de algoritmos de búsqueda en lugar de muestreo aleatorio o diseño de experimentos y finalmente, los métodos iterativos involucran un proceso recurrente de generación de subconjuntos de configuraciones y su evaluación (Huang et al., 2019). Los métodos iterativos a su vez pueden clasificarse según el mecanismo de generación de nuevas configuraciones. Algunos de los mecanismos más utilizados son: diseño experimental, optimización numérica no basada en gradiente, búsqueda heurística y optimización basada en modelos (Coy et al., 2001; Hutter et al., 2011; Riff & Montero, 2013; Bartz-Beielstein & Zaefferer, 2017).

Sin embargo, la sintonización de algoritmos de optimización no es una práctica común en

estimación de parámetros, especialmente en el caso de modelos matemáticos de procesos biotecnológicos. Esto es un grave problema considerando el teorema "non-free lunch", el cual establece que una única configuración de un optimizador no tiene la misma eficiencia para diferentes problemas (Wolpert & Macready, 1997). En este capítulo se analizarán los efectos de la sintonización y diferentes tipos de normalización de función objetivo en el desempeño de los optimizadores estocásticos enjambre de partículas, templado simulado y algoritmo genético presentes en Matlab® R2018, enfatizando en la convergencia y reproducibilidad del valor de la función objetivo alcanzada por estos optimizadores. Adicionalmente se propone una herramienta híbrida para la sintonización de algoritmos de optimización por interconexión entre Matlab® y R®.

4.3. Metodología

A continuación, se presenta la metodología propuesta en este capítulo para sintonización de algoritmos de optimización utilizados en estimación de parámetros de modelos matemáticos de procesos biotecnológicos (Figura **4-1**). La sintonización de algoritmos de optimización requiere de una selección previa del optimizador, basada en pruebas estadísticas no paramétricas (Friedman o Kruskal-Wallis y Dunn) que comparan el desempeño de diferentes algoritmos de optimización en el problema de optimización específico. Posteriormente, el optimizador seleccionado se sintoniza con el algoritmo "irace". EL proceso de sintonización requiere de la especificación del problema de optimización como un archivo de Matlab® (`Optimizador.m`) y la definición de la configuración de "irace" como archivos de texto (`parameters.txt`, `scenario.txt`).

4.3.1. Modelo matemático de bioproducción de xilitol

El xilitol es un azúcar natural de 5 carbonos que es empleado como reemplazo de edulcorantes comunes como la sacarosa, debido a su bajo contenido calórico y su protección contra la caries. Además, el xilitol presenta una amplia gama de aplicaciones en la industria farmacéutica y alimenticia como un aditivo multipropósito (Ur-Rehman et al., 2015). Sirisansaneeyakul *et al.* desarrollo una serie de 22 experimentos en modalidad batch repetidos para la producción de xilitol (Sirisansaneeyakul et al., 2013). Estos experimentos cuantificaron el crecimiento diaúxico en glucosa y xilosa del microorganismo *Candida magnoliae* su producción de xilitol durante el tiempo de fermentación. Las concentraciones de xilosa, glucosa y xilitol fueron medidas por cromatografía líquida de alta presión (HPLC por sus siglas en inglés), la concentración de biomasa fue cuantificada por gravimetría. Las condiciones de fermentación se controlaron a pH 7.0, agitación de 300 rpm, tasa de aireación de 10 vvm y 30°C.

El modelo matemático propuesto por Tochampa *et al.* (ecuaciones 4-1 - 4-7) describe la producción de xilitol en fermentación batch con crecimiento diaúxico en glucosa y xilosa para

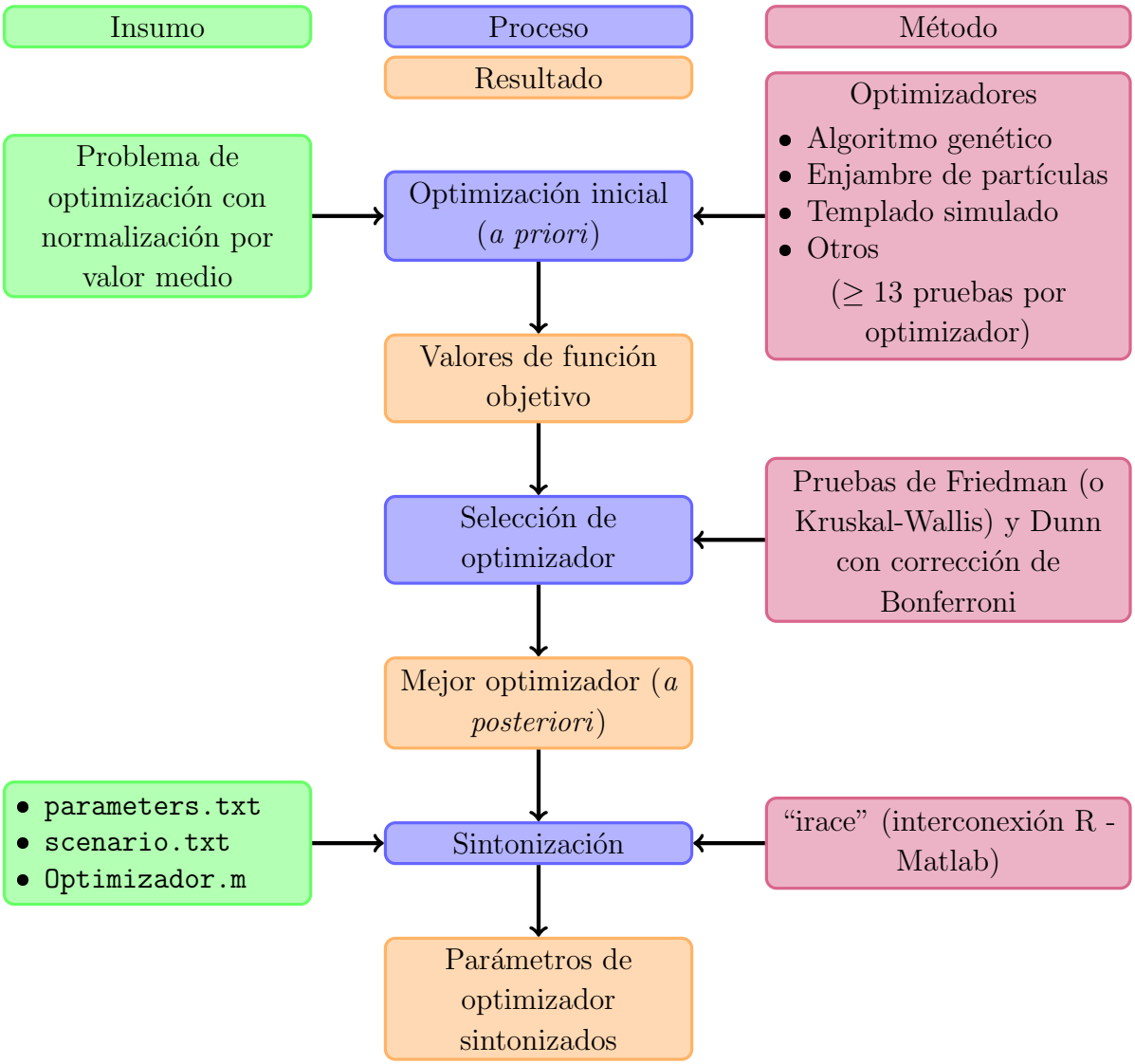


Figura 4-1: Metodología para sintonización de algoritmos de optimización utilizados en modelos de procesos biotecnológicos.

el microorganismo (Tochampa et al., 2015). Este modelo matemático está compuesto por 5 estados y 11 parámetros que considera (i) inhibición de la captación de sustratos por glucosa y xilosa, (ii) transporte de xilitol desde el interior al exterior de la célula y (iii) estequiometría de la reacción de xilosa a xilitol. Este modelo fue resuelto con el integrador "ode15s" de Matlab® R2018.

$$\frac{dC_X}{dt} = \mu C_X \tag{4-1}$$

$$\frac{dC_{glu}}{dt} = -\left[q_{glu}^{max}\frac{C_{glu}}{C_{glu} + K_{S,glu}\left(1 + \frac{C_{xil}}{K_{i,xil}}\right)}\right] C_X \tag{4-2}$$

$$\frac{dC_{xil}}{dt} = -\left[q_{xil}^{max}\frac{C_{xil}}{C_{xil} + K_{S,xil}\left(1 + \frac{C_{glu}}{K_{i,glu}}\right)}\right] C_X \tag{4-3}$$

$$\frac{dC_{xit}^{in}}{dt} = \rho_X(r_{f,xit} - r_{u,xit} - r_{t,xit}) - \mu C_{xit}^{in} \tag{4-4}$$

$$\frac{dC_{xit}^{ex}}{dt} = r_{t,xit}C_X \tag{4-5}$$

$$\mu = \mu_{glu}^{max}\frac{C_{glu}}{K_{S,glu} + C_{glu}} + \mu_{xit}^{max}\frac{C_{xit}^{in}}{K_{S,xit} + C_{xit}^{in}}\frac{K_r}{K_r + C_{glu}} \tag{4-6}$$

$$r_{t,xit} = 3.6\text{x}10^6 P_{xit} a_{cell}(C_{xit}^{in} - C_{xit}^{ex}) \tag{4-7}$$

4.3.2. Función objetivo

La función objetivo cuantifica la desviación entre la repuesta del modelo y datos experimentales. Entre las funciones objetivo más comunes se encuentra la función objetivo de mínimos cuadrados, la cual presenta propiedades estadísticas deseables para estimación de parámetros como generar estimados de máxima verosimilitud, eficientes y no sesgados (Quinn & Keough, 2002). Sin embargo, las diferencias en escala de los datos experimentales afectan el proceso de optimización llevando a un ajuste inequitativo de los datos experimentales (Quinn & Keough, 2002; Charaniya et al., 2008; Ji, 2012). Para mitigar este problema, términos de normalización son comúnmente incluidos en la función objetivo.

Entre los términos de normalización utilizados se encuentran el valor máximo o medio de la variable experimental o la desviación estándar de los experimentos. Adicionalmente, es posible incluir en la función de mínimos cuadrados múltiples experimentos con diferente cantidad de datos experimentales. La ecuación 4-8 corresponde a la función objetivo de mínimos cuadrados con factor de normalización ω (máximo, mínimo o media del estado experimental) y considera diferencias en el número de experimentos N_{exp}, estados experimentales N_{var} y número de puntos experimentales N_{obs}. Adicionalmente, $\mathbf{x}_{j,i,k}$ corresponde a la salida del modelo matemático, $\mathbf{y}_{j,i,k}$ a datos experimentales y $\varphi(\boldsymbol{\theta})$ valor de función objetivo.

$$\varphi(\boldsymbol{\theta}) = \sum_{k=1}^{N_{exp}}\sum_{j=1}^{N_{var}}\sum_{i=1}^{N_{exp}}\frac{\left(\mathbf{y}_{j,i,k} - \mathbf{x}_{j,i,k}\right)^2}{\omega_{j,k}} \tag{4-8}$$

4.3.3. Algoritmos de optimización

Los algoritmos de optimización estocásticos, o también llamados metaheurísticas, exploran el espacio de búsqueda de un problema de optimización a través de reglas derivadas de la naturaleza (Mirjalili et al., 2020). Este tipo de optimizadores pueden ser catalogados de manera general en algoritmos de ascenso de colina, inteligencia de enjambre y algoritmos evolutivos. Algunas de las metaheurísticas más utilizadas pertenecientes a estas categorías son templado simulado, optimización por enjambre de partículas y algoritmo genético, respectivamente (Ezugwu et al., 2021). De forma particular, se ha reportado la utilidad de estas metaheurísticas en estimación de parámetros de modelos de procesos biotecnológicos (Moles et al., 2003; Banga et al., 2004).

- Templado simulado

  El optimizador de templado simulado (simulated annealing o SA por sus siglas en inglés) es una metaheurística que se basa en un enfriamiento controlado de un grupo de átomos para obtener un cristal perfecto, es decir, con la menor energía (Dowsland & Thompson, 2012). Este algoritmo de optimización está controlado por los parámetros presentados en la Tabla **4-1**, cuyos rangos son descritos en el *global optimization toolbox* de Matlab® R2018. Este optimizador posee cuatro parámetros siendo uno de ellos racional o continuo, uno entero y dos de tipo categórico. La descripción detallada de estos parámetros se da en el apéndice B.3.1.

Tabla 4-1: Parámetros y rangos del optimizador templado simulado.

Parámetro	Tipo	Rango/categoría
InitialTemperature	Racional	$1x10^{-6}$ - 500
ReannealInterval	Entero	1 - 200
TemperatureFcn	Categórico	temperatureexp temperaturefast temperatureboltz
AnnealingFcn	Categórico	annealingfast annealingboltz

- Enjambre de partículas

  El optimizador de enjambre de partículas (Particle Swarm Optimization o PSO por sus siglas en inglés) es una metaheurística que se basa en el comportamiento social de un grupo (enjambre) de agentes (partículas) que buscan la localización del mejor recurso (Poli et al., 2007). Este algoritmo está controlado por los parámetros presentados en la Tabla **4-2**, cuyos rangos son descritos en el *global optimization toolbox* de Matlab® R2018. El algoritmo enjambre de partículas posee seis parámetros, cinco de racional o

continuo y uno de tipo entero. La descripción detallada de estos parámetros se da en el apéndice B.3.2.

Tabla 4-2: Parámetros y rangos del optimizador enjambre de partículas.

Parámetro	Tipo	Rango
IntertiaRangelb	Racional	0.1 - 1.1
InertiaRangeub	Racional	0.1 - 1.1
MinNeighborsFraction	Racional	0 - 1
SelfAdjustmentWeight	Racional	0.5 - 2.5
SocialAdjustmentweight	Racional	0.5 - 2.5
SwarmSize	Entero	2 - 200

- Algoritmo genético

El algoritmo genético (genetic algorithm o GA por sus siglas en inglés) es una metaheurística que se basa en la evolución de una población de individuos sometidos a una presión ambiental que se ven modificados por fuerzas evolutivas (mutación, entrecruzamiento, inmigración, emigración y selección) (Sakawa, 2012). Este optimizador estocástico presenta los parámetros mostrados en la Tabla **4-3**. Los rangos de los parámetros son recomendados en la documentación del *global optimization toolbox* de Matlab® R2018. El optimizador algoritmo genético presenta 14 parámetros en total, siendo tres de ellos de tipo entero, siete de tipo continuo o racional y cuatro de tipo categórico. Así mismo, cuenta con parámetros adicionales dependiendo de sí se seleccionan ciertas opciones de los parámetros categóricos. Entre estos parámetros "dependientes" se tienen cinco de tipo continuo o racional y uno de tipo entero. La descripción detallada de estos parámetros se da en el apéndice B.3.3.

La cantidad de parámetros a sintonizar en los algoritmos de optimización mencionados varía de cuatro a catorce, que además, abarcan variables tanto de tipo cuantitativo como cualitativo. Lo anterior conllevaría a una cantidad elevada de combinaciones en caso de utilizar un método simple de sintonización. Los métodos de alto nivel presentan el inconveniente de usar una metaheurística no sintonizada para sintonizar otra metaheurística, lo cual puede derivar en altos costos computacionales con un pobre desempeño. Los métodos iterativos de sintonización proveen un resultado estadísticamente robusto, con un uso eficiente de tiempo y recursos computacionales al descartar prematuramente configuraciones con bajo desempeño.

Tabla 4-3: Parámetros y rangos del optimizador algoritmo genético.

Parámetro	Tipo	Rango/categorias	Parámetros adicionales
PopulationSize	Entero	1 - 400	No aplica
CrossoverFraction	Racional	0 - 1	No aplica
MigrationDirection	Categórico	forward, both	No aplica
MigrationInterval	Entero	1 - 20	No aplica
MigrationFraction	Racional	0.001 - 1	No aplica
SelectionFcn	Categórico	selectionstochunif, selectionremainder, selectionuniform, selecionroulette, selectiontournament	No aplica
CrossoverFcn	Categórico	crossoverscattered, corssoversinglepoint, corssoverintermediate, crossoverheuristic, crossoverarithmetic	No aplica
MutationFcn	Categórico	mutationgaussian, mutationuniform	No aplica
Shrink	Racional	-2 - 2	mutationgaussian
Scale	Racional	0.1 - 10	mutationgaussian
Mrate	Racional	0.01 - 0.5	mutationuniform
Tsize	Entero	2 - 15	selectiontournament
CIratio	Racional	0 - 1	crossoverintermediate
CHratio	Racional	0 - 1	crossoverheuristic

4.3.4. Herramienta híbrida para sintonización de algoritmos de optimización

La herramienta híbrida de sintonización de algoritmos de optimización se desarrolló mediante interconexión entre los software Matlab® y R® a través del paquete "R.matlab". Este paquete permite el envío de comandos desde R® a Matlab® y la recuperación de variables Matlab® nuevamente a R®. En este sentido, R® es el lenguaje maestro y Matlab® el esclavo. La sintonización de los algoritmos de optimización fue realizada con el paquete "irace". El paquete "irace" desarrollado por López *et al.* en 2016 permite realizar un análisis con el método *racing* (ver apéndice B.1) de forma iterativa y automática para la sintonización de algoritmos computacionales (López-Ibáñez et al., 2016).

El algoritmo de la herramienta híbrida se muestra en la Figura **4-2**. En su primera iteración "irace" genera configuraciones aleatorias en R® y las evalúa en Matlab®, este proceso

lleva a la construcción de funciones de distribución de probabilidad para los parámetros del optimizador. En iteraciones posteriores, nuevas configuraciones son muestreadas a partir de las distribuciones de probabilidad generadas. De esta forma, configuraciones ineficientes son descartadas rápidamente a través de pruebas estadísticas y nuevas configuraciones promisorias son generadas con mayor probabilidad en cada iteración, lo que se traduce en menor esfuerzo computacional.

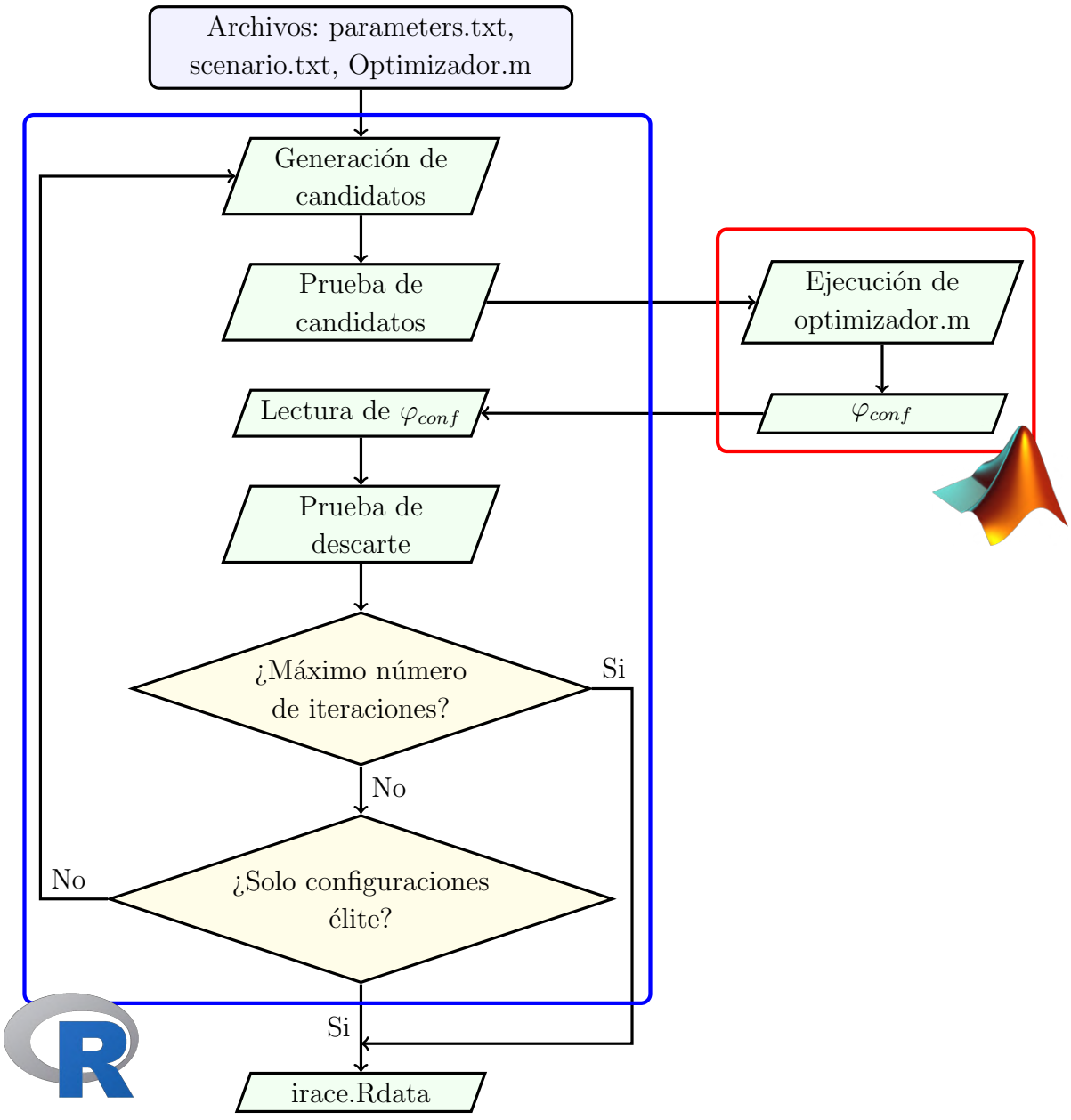


Figura 4-2: Herramienta híbrida para sintonización de algoritmos de optimización por interconexión R-Matlab.

Tres archivos de entrada son requeridos para el algoritmo "irace": `parameters.txt`,`scenario.txt` y `Optimizador.m`. `parameters.txt` especifica los nombres, tipos (e.g. entero, categórico, continuo) y rangos de los parámetros del optimizador. `scenario.txt` especifica opciones internas del algoritmo "irace". Por último, `Optimizador.m` es un código de Matlab que especifica el problema de optimización cuya entrada corresponde a la configuración del optimizador generada por "irace" y su salida es el valor de función objetivo. `irace.Rdata` es la salida del algoritmo "irace" y contiene información del proceso de sintonización como configuraciones generadas, valores de función objetivo calculados, configuraciones élite, número de iteraciones de *racing*, entre otros.

"irace" presenta mejoras como muestro de configuraciones desde distribuciones normales truncadas, estrategia de reinicio para prevenir convergencia prematura y un *racing* elitista que preserva las mejores configuraciones generadas hasta el momento, lo que asegura equidad en las comparaciones entre configuraciones. Además, cada configuración promisoria es probada en diferentes instancias (semillas aleatorias) de un generador de números aleatorios, lo que garantiza robustez de la configuración a diferentes procesos estocásticos (L'Ecuyer, 2012). Una descripción general del funcionamiento del paquete "irace" se presenta en el apéndice B.2.

4.3.5. Diseño experimental basado en simulación para evaluación del desempeño de algoritmos de optimización

Se realizó en un diseño experimental de bloques, con tres optimizadores (algoritmo genético, enjambre de partículas y templado simulado) y tres normalizaciones de función objetivo (valores medio, máximo y mínimo por estado experimental), lo que resulta en nueve combinaciones que fueron sintonizadas con la herramienta híbrida. Para garantizar reproducibilidad y precisión, la mejor configuración sintonizada por combinación de optimizador y factor de normalización fue probada 45 veces. Para la comparación de los resultados de optimización, los valores φ_{conf} calculados con máximo y mínimo fueron traducidos a su equivalente en factor de normalización de media. Esto permite una escala común con mayor interpretabilidad, dado que su magnitud es cercana a aquella de los datos experimentales.

La comparación llamada "A" presenta la diferencia general entre los factores de normalización (grupos) bajo diferentes optimizadores (bloques). La comparación "B" presenta la diferencia general entre los optimizadores (grupos) bajo diferentes factores de normalización (bloques). Los resultados de optimización fueron comparados con la prueba estadística no paramétrica de Friendman y en caso de encontrar diferencias significativas, la prueba de Dunn para comparaciones múltiples con corrección de Bonferroni fue realizada para especificar diferencias entre grupos (Fahoome & Sawilowsky, 2002; Dinno, 2017). Para determinar si

existen diferencias en el desempeño entre configuraciones sintonizadas y por defecto (Tabla **4-4**), una prueba no paramétrica de Friedman adicional fue hecha.

Tabla 4-4: Configuraciones por defecto para los optimizadores presentes en el *global optimization toolbox* de Matlab® R2018.

Algoritmo genético		Enjambre de partículas	
CrossoverFcn	crossoverscattered	InertiaRangelb	0.1
CrossoverFraction	0.8	InertiaRangeub	1.1
MigrationDirection	forward	MinNeighborsFraction	0.25
MigrationFraction	0.2	SelfAdjustmentWeight	1.49
MigrationInterval	20	SocialAdjustmentWeight	1.49
MutationFcn	mutationgaussian	SwarmSize	100
Scale	1	Templado simulado	
Shrink	1	AnnealingFcn	annealingfast
PopulationSize	200	InitialTemperature	100
SelectionFcn	selectionstochunif	ReannealInterval	100
-	-	TemperatureFcn	temperatureexp

4.4. Resultados

A continuación, se presentan la evaluación sistemática del proceso de sintonización para los algoritmos de optimización. Primero, se presentan los resultados generales del proceso de sintonización para los tres optimizadores y cada factor de normalización. Además, se muestran los resultados concernientes al comportamiento exploratorio de cada optimizador sintonizado. Posteriormente, el desempeño en términos de precisión y reproducibilidad del valor de función objetivo para los optimizadores sintonizados y por defecto es analizado. Finalmente, se presentan las comparaciones estadísticas del desempeño entre combinaciones sintonizadas de optimizador y factor de normalización.

4.4.1. Sintonización de algoritmos de optimización global

Resultados generales

Los resultados generales del proceso de sintonización para los optimizadores SA, GA y PSO se muestran en la Tabla **4-5**. Los tres optimizadores presentan resultados similares tanto para el número de configuraciones generadas sin importar el factor de normalización como el φ_{conf} calculado más bajo. Esto indica que por lo menos una configuración del optimizador pudo alcanzar el mínimo global de la función objetivo de mínimos cuadrados en todos

los casos con el factor de normalización especificado. Sin embargo, el φ_{conf} más alto difiere por varios órdenes de magnitud entre resultados con el mismo factor de normalización, lo cual puede ser explicado por la generación de configuraciones que llevan a combinaciones de parámetros no convergentes. El optimizador PSO presenta el mejor comportamiento de búsqueda con un menor valor máximo de φ_{conf} para los factores de normalización mínimo (9219) y máximo (37.75) en comparación con GA (12118, 153884) y SA (1.57x10^9, 1024).

Respecto al esfuerzo computacional, la sintonización del optimizador GA es significativamente más lenta que en los demás optimizadores ($\approx 300\,\%$), lo cual es de manera más probable, debido al alto número de parámetros a ser sintonizados (14) comparado con PSO (6) y SA (4). En términos generales, la tarea que demanda mayor esfuerzo computacional (consume más tiempo) en el proceso de sintonización corresponde a la solución del problema de optimización con cada configuración de hiperparámetros generada por el sintonizador. Para el caso de estudio analizado, el optimizador GA presenta las configuraciones más ineficientes comparado con SA y PSO, lo que se ve reflejado en su elevado esfuerzo computacional y mayor número de configuraciones generadas. Las diferencias mostradas en el proceso de sintonización reflejan el comportamiento inherente de estos algoritmos de optimización.

Tabla 4-5: Resultados generales para el proceso de sintonización de los algoritmo de optimización SA, PSO y GA.

Algoritmo de optimización		Factor de normalización		
		Mínimo	Máximo	Media
Templado simulado	Configuraciones generadas	129	135	119
	φ_{conf} máximo	153884	1024	20.49
	φ_{conf} mínimo	346	0.274	1.07
	Mejor configuración	80	126	105
	Esfuerzo computacional (s)	85284	83448	93852
Enjambre de partículas	Configuraciones generadas	140	138	117
	φ_{conf} máximo	9219	37.75	41.31
	φ_{conf} mínimo	376	0.018	1.31
	Mejor configuración	119	129	96
	Esfuerzo computacional (s)	49860	62316	54504
Algoritmo genético	Configuraciones generadas	153	161	158
	φ_{conf} máximo	12118	1.57x10^9	4334
	φ_{conf} mínimo	351.9	0.018	1.04
	Mejor configuración	113	128	92
	Esfuerzo computacional (s)	277x10^3	244x10^3	235x10^3

- Templado simulado

El conjunto completo de resultados específicos para el optimizador SA con cada factor de normalización se presenta en el Apéndice B.4.1. Para cada factor de normalización, la misma tendencia aproximada aparece para la dispersión de φ_{conf} contra el número de configuraciones generadas usando los factores de normalización mínimo (Figura **4-3**) y media (Figura **B-9**). Al inicio del proceso o de sintonización, las configuraciones exhiben diagramas de caja con dispersiones substanciales en sus rangos intercuartil (RIC) con altos valores de función objetivo, lo que implica configuraciones ineficientes que presentan convergencia lenta o prematura y/o inhabilidad para escapar de mínimos locales. Lo anterior es debido a un comportamiento inadecuado del optimizador debido a una relación desfavorable de exploración (búsqueda global) y explotación (búsqueda local) (Chen et al., 2009). Además, esto implica que la configuración por defecto del optimizador puede no ser apropiada para este problema específico y el proceso de sintonización no es trivial.

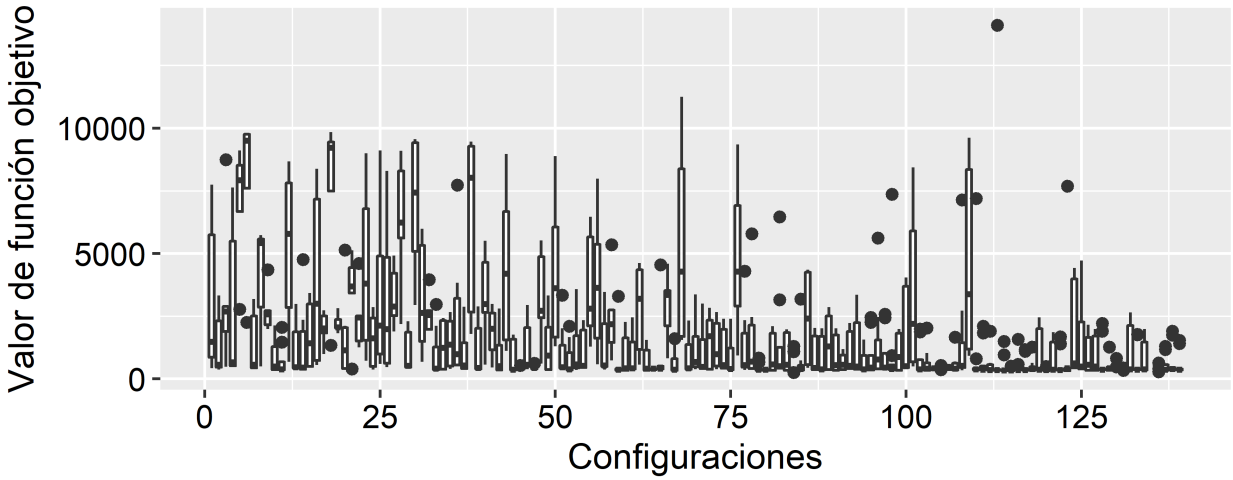


Figura 4-3: Dispersión de valores de función objetivo para el optimizador SA con factor de normalización mínimo.

A medida que el número de configuraciones generadas aumenta en la Figura **4-3**, φ_{conf} y RIC se reducen a un mínimo alrededor de la centésima configuración con el mismo comportamiento en configuraciones posteriores. Esto muestra que el proceso de sintonización ha generado conjuntos de configuraciones que presentan propiedades exploratorias eficientes y son robustas a puntos de inicio aleatorios. Esto implica que las configuraciones pueden efectivamente escapar de mínimos locales, explorar todo el espacio de búsqueda y converger al mínimo global sin importar el punto inicial seleccionado (Jasrasaria & Pyzer-Knapp, 2018). En el caso del factor de normalización máximo (Figura **4-4**), los resultados para la dispersión de φ_{conf} contra el número de configuraciones generadas muestra un patrón interesante en donde solo algunas configuraciones exhiben dispersiones considerables. Esto sugiere que configuraciones eficientes son encontradas desde el inicio del proceso de sintonización en contraste con los factores de normalización mínimo y media.

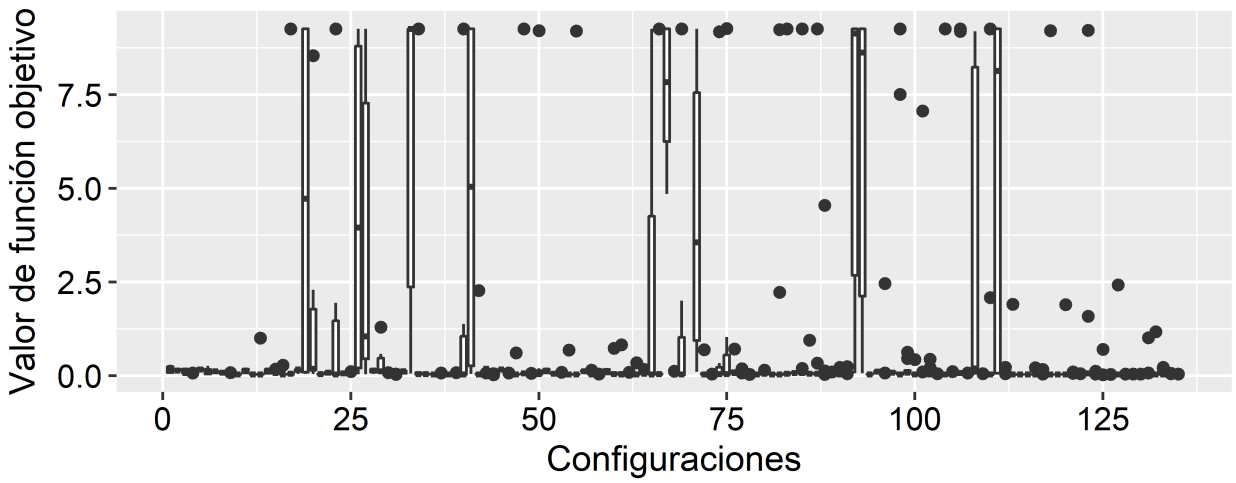


Figura 4-4: Dispersión de valores de función objetivo para el optimizador SA con factor de normalización máximo.

No existe una única configuración de optimizador que presenta una relación exploración-explotación para una combinación dada de algoritmo y problema de optimización. Esto se confirma por la presencia de múltiples configuraciones élite con desempeño similar para el algoritmo SA y su combinación con factores de normalización media (Figura **4-5**), mínimo y máximo (Figuras **B-6** y **B-7**). Sin embargo, en cada caso la configuración mejor sintonizada corresponde a aquella con menor dispersión de φ_{conf} y mediana. La mejor configuración calculada para el optimizador SA se muestra en la Tabla **4-6**. La diferencia entre las configuraciones sintonizadas sugiere un efecto del factor de normalización en el comportamiento del optimizador.

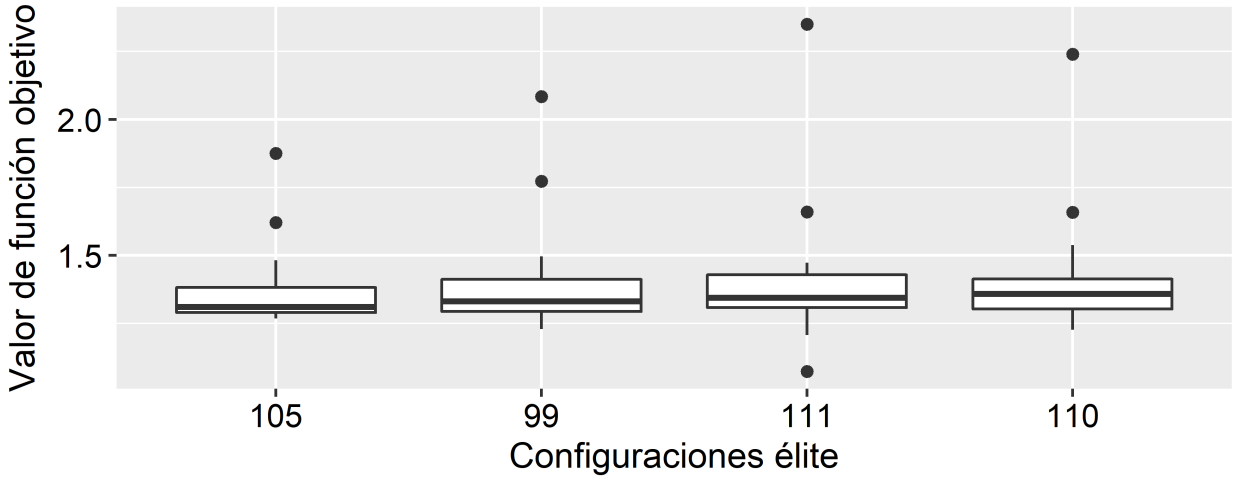


Figura 4-5: Dispersión de valores de función objetivo para el optimizador SA con factor de normalización media.

La evolución de los valores de los parámetros internos del optimizador a lo largo del proceso de sintonización también provee información relevante. Para el algoritmo SA, los parámetros `InitialTemperature`, `ReannealInterval`, `TemperatureFcn` y `AnnealingFcn` presentan un comportamiento diferente dependiendo del factor de normalización. Los factores de normalización mínimo (Figura **4-6**) y media (Figura **B-10**) comparten una forma unimodal en las densidades de probabilidad de `InitialTemperature` y `ReannealInterval`, con una moda de valor bajo para el factor de normalización mínimo y moda de valor intermedio para el factor de normalización media. Esta diferencia puede deberse a la interacción con el parámetro `TemperatureFcn` y su categoría "Boltz", siendo esta categoría predominante en la normalización con valor mínimo y la categoría "Fast" en la normalización con media. Lo anterior implica que un descenso rápido en la temperatura del algoritmo SA (controlada por `TemperatureFcn`) es compensada con un recocido retrasado (controlado por `ReannealInterval`), el cual controla la relación exploración-explotación.

Tabla 4-6: Mejores configuraciones del algoritmo SA.

Parámetro	Factor de normalización			
	Mínimo		Medio	Máximo
Configuración	80		105	126
`InitialTemperature`	12.9808		5.5345	207.44
`ReannealInterval`	66		107	168
`TemperatureFcn`	Boltz	Fast	Fast	
`AnnealingFcn`	Boltz	Boltz	Boltz	

Dentro de mecanismo interno del optimizador SA, altas temperaturas están asociadas con exploración, por el contrario, bajas temperaturas lo están con explotación. Por otra parte, el parámetro `ReannealInterval` está asociado con la posibilidad de realizar una nueva búsqueda global, dado que este eleva la temperatura a su valor inicial (Siddique & Adeli, 2016). El proceso de sintonización para SA con factor de normalización máximo (Figura **B-8**) tiende a ejecutar más búsquedas globales (alta exploración y baja exploración), por otra parte, el proceso de sintonización para este algoritmo con factor de normalización media tiende a ejecutar menor cantidad de búsquedas globales con búsqueda local extendida (menor exploración y mayor explotación). La categoría "Boltz" para el parámetro `AnnealingFcn` es predominante con todos los factores de normalización.

El proceso de sintonización para el algoritmo SA con factor de normalización máximo presenta una interacción más compleja entre sus parámetros con forma bimodal en la densidad de probabilidad para `InitialTemperature` y `ReannealInterval`. Sin embargo, las categorías predominantes para `TemperatureFcn` y `AnnealingFcn` son "Fast" y "Boltz", respectivamente. Como la moda con mayor probabilidad para `ReannealInterval`, es posible afirmar que el proceso de sintonización para el optimizador SA con factor de normalización máximo es si-

milar a aquel con factor de normalización media, usando baja exploración y alta explotación. La categoría "Boltz" en el parámetro `AnnealingFcn` para todos los factores de normalización implica que el algoritmo SA tiene mejor desempeño con pasos cortos en las variables de optimización para este caso.

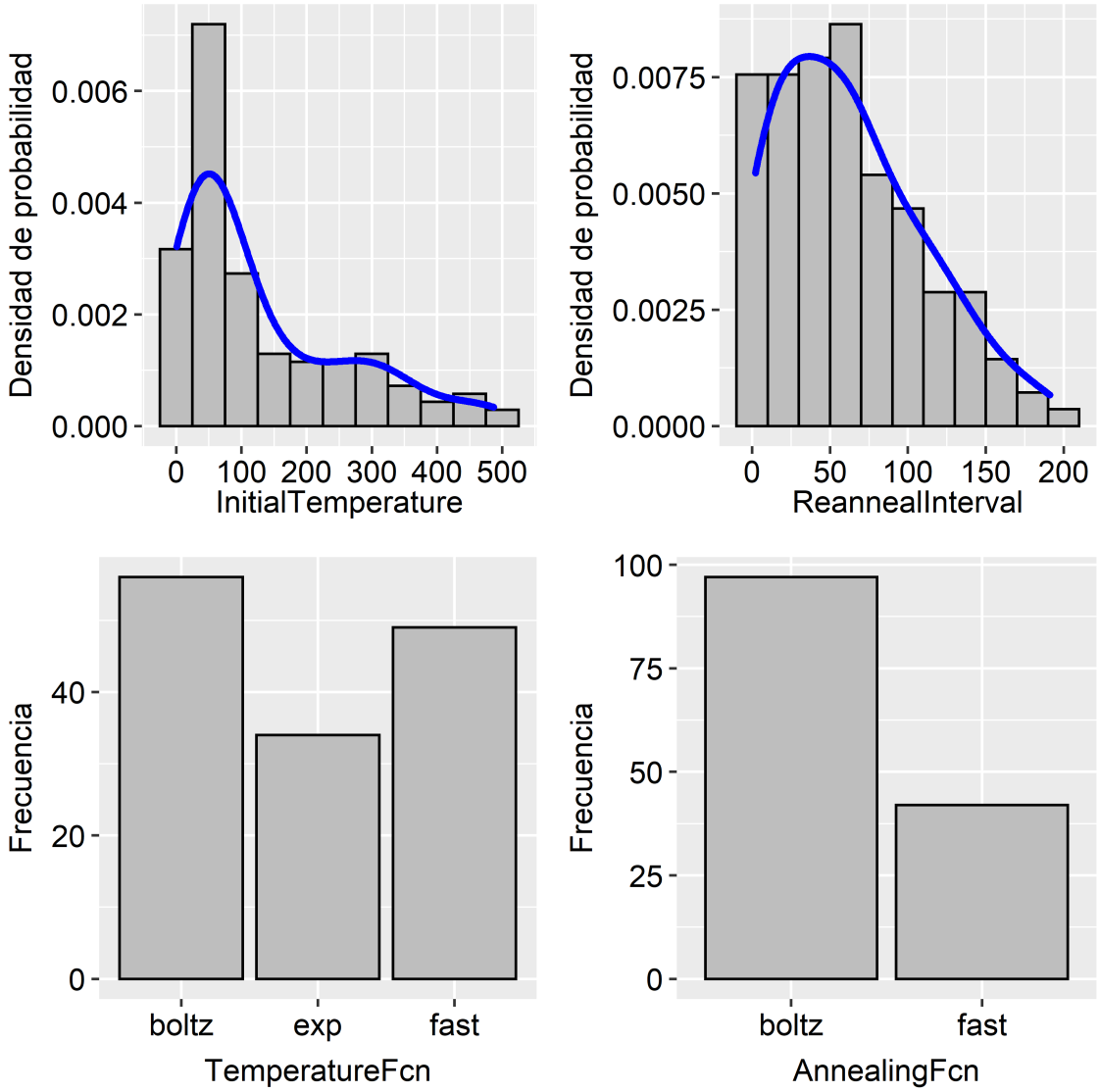


Figura 4-6: Diagrama de frecuencia para los parámetros de templado simulado con normalización por valor mínimo para `InitialTemperature`, `RannealInterval`, `TemperatureFcn` y `AnnealingFcn`.

- Enjambre de partículas

El proceso de sintonización del algoritmo PSO produce un resultado similar para la dispersión de φ_{conf} contra el número de configuraciones generadas empleando los tres factores de

normalización (Figuras **B-11**, **B-14** y **B-17**). En todos los casos, cerca de la configuración número 75, la dispersión de φ_{conf} colapsa para la mayor parte de las configuraciones siguientes. Esto sugiere que configuraciones con una relación exploración-explotación han sido encontradas. Sin embargo, en el caso del factor de normalización máximo, configuraciones eficientes fueron encontradas desde el inicio del proceso de sintonización.

Para cada factor de normalización, en el proceso de sintonización se calcularon cuatro configuraciones élite con desempeño similar (Figuras **B-12**, **B-15**y **B-18**). Las mejores configuraciones élite para el proceso de sintonización del algoritmo PSO y los tres factores de normalización se muestra en la Tabla **4-7**. Las diferencias entre los valores de los parámetros del optimizador indican que el factor de normalización tiene una fuerte influencia en el comportamiento exploratorio del algoritmo de optimización.

Tabla 4-7: Mejores configuraciones del algoritmo PSO

Parámetro	Factor de normalización		
	Mínimo	Media	Máximo
Configuración	119	96	129
`InertiaRangelb`	0.6414	0.2517	0.6837
`InertiaRangeub`	0.6762	0.5043	0.7152
`MinNeighborsFraction`	0.6508	0.3559	0.7152
`SelfAdjustmentWeight`	1.0179	1.5869	1.0180
`SocialAdjustmentWeight`	1.7858	1.4422	1.8206
`SwarmSize`	144	185	174

La diferencia en los valores de los parámetros sintonizados para el optimizador PSO provee información acerca de la influencia del factor de normalización en la relación exploración-explotación en el espacio de búsqueda. Los diagramas de frecuencia paramétrica de los seis parámetros del algoritmo PSO: `InertiaRangelb`, `InertiaRangeub`, `MinNeighborsFraction`, `SelfAdjustmentWeight`, `SocialAdjusmentWeight` y `SwarmSize` para cada factor de normalización se muestran en las Figuras **B-13**, **B-16** y **B-19**.

Los parámetros del optimizador PSO muestran un comportamiento complejo para todos los casos con densidades de probabilidad que varían en forma desde unimodal a multimodal, especialmente en `SelfAdjustmentWeight` y `SocialAdjustmentWeight`. Para el factor de normalización mínimo, los hiperparámetros `InertiaRangelb` e `InertiaRangeub` exhiben valores similares, `MinNeighborsFraction` tiende a un valor elevado y `SelfAdjusmentWeight` tiende a un valor bajo, mientras `SocialAdjusmentWeight` presenta un rango uniforme de valores mayores que `SocialAdjustmentWeight`. Lo anterior implica que bajo el factor de normalización mínimo, el algoritmo PSO evidencia baja exploración con alta explotación.

Las partículas del enjambre muestran un alto sesgo hacia la misma posición en la iteración actual (`InertiaRangelb` e `InertiaRangeub`), lo que se traduce en pasos de exploración cortos. Además, las partículas intercambian información con una fracción `MinNeighborsFraction` mayoritaria de todo el enjambre y presentan inclinación a seguir la partícula con la mejor posición actual (`SocialAdjustmentWeight`), en lugar de realizar una búsqueda individual exhaustiva (`SelfAdjustmentWeight`). Esto se traduce en alta exploración (Wang et al., 2018a).

El proceso de búsqueda del algoritmo PSO con factor de normalización media es análogo al de factor de normalización mínimo, excepto que las partículas tienden a realizar pasos más anchos desde su posición actual y la comunicación entre las partículas está significativamente reducida, lo que sugiere menor explotación. El comportamiento del algoritmo de optimización PSO bajo factor de normalización máximo es de nuevo similar a los resultados obtenidos con los factores de normalización media y mínimo. Sin embargo, existe una tendencia mayor a mantener la posición actual, lo cual indica mayor explotación. El parámetro `SwarmSize` tiene consistencia entre los tres factores de normalización, con forma unimodal en su densidad de probabilidad y una moda alrededor de 150 partículas. Esto significa que en promedio, la información individual y colectiva de 150 partículas es requerida para explorar eficientemente el espacio de búsqueda.

- Algoritmo genético

Las dispersiones de φ_{conf} contra el número de configuraciones generadas con los tres factores de normalización para el optimizador GA son mostradas en las Figuras **B-20**, **B-25** y **B-30**. Los tres factores de normalización presenta una tendencia similar con el colapso de las dispersiones de φ_{conf} a un valor mínimo alrededor de la configuración generada número 100. Esto significa que una relación de exploración-explotación competente ha sido encontrada para esta combinación de algoritmo y problema de optimización.

El proceso de sintonización del optimizador Ga ha calculado cinco configuraciones élite para cada factor de normalización y sus diagramas de caja se muestran en las Figuras **B-21**, **B-26** y **B-31**. La dispersión de φ_{conf} para las configuraciones élite es similar para cada factor de normalización. Esto implica que la existencia de múltiples combinaciones entre los parámetros del optimizador que dan como resultado una relación de exploración-explotación adecuada. Las características de las mejores configuraciones para cada factor de normalización se muestra en la Tabla **4-8**. Los tres factores de normalización muestran resultados idénticos para los operadores de evolución, sin embargo, los valores de los parámetros internos de estos operadores presentan una diferencia significativa entre los factores de normalización mínimo y ambos, media y máximo, lo que indica influencia del factor de normalización en el proceso de sintonización.

Tabla 4-8: Mejores configuraciones del optimizador GA.

Parámetro	Factor de normalización		
	Mínimo	Media	Máximo
`PopulationSize`	362	373	335
`MigrationDirection`	"Forward"	"Forward"	"Forward"
`MigrationFraction`	0.1139	0.7105	0.8389
`MigrationInterval`	11	16	15
`SelectionRemainder`	"Remainder"	"Roulette"	"Stochunif"
`CrossoverFraction`	0.5535	0.7680	0.7846
`CrossoverFcn`	"Heuristic"	"Heuristic"	"Heuristic"
`CHratio`	0.6722	0.3156	0.3189
`MutationFcn`	"Uniform"	"Uniform"	"Uniform"
`Mrate`	0.1467	0.2462	0.2316

Debido a la cantidad considerable de parámetros del optimizador GA, interacciones complejas se forman en los diagramas de frecuencia mostrados en las Figuras **B-22** a **B-24**, **B-27** a **B-29** y **B-32** a **B-34**. Para los tres factores de normalización los parámetros `CrossoverFcn`, `MutationFcn` y `MigrationDirection` comparten los operadores "Heuristic", "Uniform" y "Forward", respectivamente, como su categoría más frecuente. El operador "Huristic" genera un porcentaje `CrossoverFraction` de nuevas soluciones promisorias por combinación lineal entre los "individuos" con mejor y peor desempeño en la generación previa. El operador "Uniform" selecciona al azar una variable de un "individuo" con una probabilidad `Mrate` y cambia su valor entre limites especificados. `MigrationDirection` especifica la dirección en la cual un porcentaje `MigrationFraction` de "individuos" de la generación previa son introducidos una cantidad `MigrationInterval` de generaciones adelante.

El proceso de sintonización del optimizador GA con los tres factores de normalización presenta diferentes categorías para `SelectionFcn`, el cual controla la selección de configuraciones parentales sobre las cuales la operación de entrecruzamiento va a ser realizada. Primero, los operadores de entrecruzamiento y selección están enfocados en explotación por combinación de información presente en los "individuos"de la generación actual. Por otra parte, los operadores de migración y mutación están enfocados en la exploración al introducir nueva información acerca del espacio de búsqueda en los "individuos" de la generación actual (Lim et al., 2017). En general, el optimizador GA presenta una relación exploración- explotación similar para los tres factores de normalización, con diferencias en la explotación debido al operador de selección. No obstante, las diferencias entre parámetros con dominio discreto o continuo como `Mrate` o `MigrationInterval` implican influencia del factor de normalización en el proceso de sintonización.

4.4.2. Desempeño de algoritmos de optimización sintonizados

a) Configuraciones por defecto

Las configuraciones por defecto utilizadas en Matlab® R2018 se muestran en la Tabla **4-4**. En comparación con las contrafiguras sintonizadas bajo los tres factores de normalización (Tablas **4-6**, **4-7** y **4-8**), las configuraciones por defecto muestran diferencias significativas tanto en los parámetros categóricos como continuos. Esto es una indicación temprana para esperar un pobre desempeño de la configuración con defecto para el problema de estimación de parámetros investigado.

Los resultados de la dispersión de φ_{conf} y la densidad de probabilidad para 4 optimizaciones con las configuraciones por defecto y factor de normalización media para los tres algoritmos de optimización son mostrados en la Figura **4-7**. En comparación con un valor φ_{conf} igual a 1.04 (mejor valor encontrado con el factor de normalización media en Tabla **4-5**), las figuras muestran un bajo desempeño para las configuraciones por defecto. El optimizador GA se quedó atascado consistentemente en un mínimo local, en contraste, los optimizadores SA y PSO pueden escapar de este mínimo local y encontrar la región del óptimo global pero con altas dispersiones. Esto implica que por lo menos los algoritmos PSO y SA pueden encontrar el mínimo global con sus configuraciones por defecto. Sin embargo, la relación exploración-explotación está pobremente definida por defecto.

b) Configuraciones sintonizadas con normalización de valor mínimo

El resultado para la dispersión de φ_{conf} y su densidad de probabilidad para los algoritmos sintonizados con factor de normalización mínimo se muestran en la Figura **B-35**. Los optimizadores SA y PSO presentan la dispersión de φ_{conf} estrecha, con pocos puntos atípicos y mediana alrededor de 400. En comparación con un valor φ_{conf} igual a 346 (mejor valor encontrado con factor de normalización mínimo en Tabla **4-5**), existe una mejora significativa con respecto a las configuraciones por defecto tanto en precisión como en reproducibilidad para estos optimizadores. El optimizador GA presenta una dispersión más amplia con una mediana de φ_{conf} alrededor de 600. Por lo tanto, el proceso de sintonización tiene un efecto significativo en la relación exploración-explotación en los optimizadores PSO y SA, sin embargo, este no fue el caso para GA en el cual dicha relación se inclina hacia explotación.

c) Configuraciones sintonizadas con normalización de valor máximo

Los resultados de la dispersión de φ_{conf} y densidad de probabilidad para los algoritmos de optimización sintonizados con factor de normalización máximo se muestran en la figura **B-36**. El optimizador GA presenta la dispersión de φ_{conf} más estrecha, no obstante, su mediana

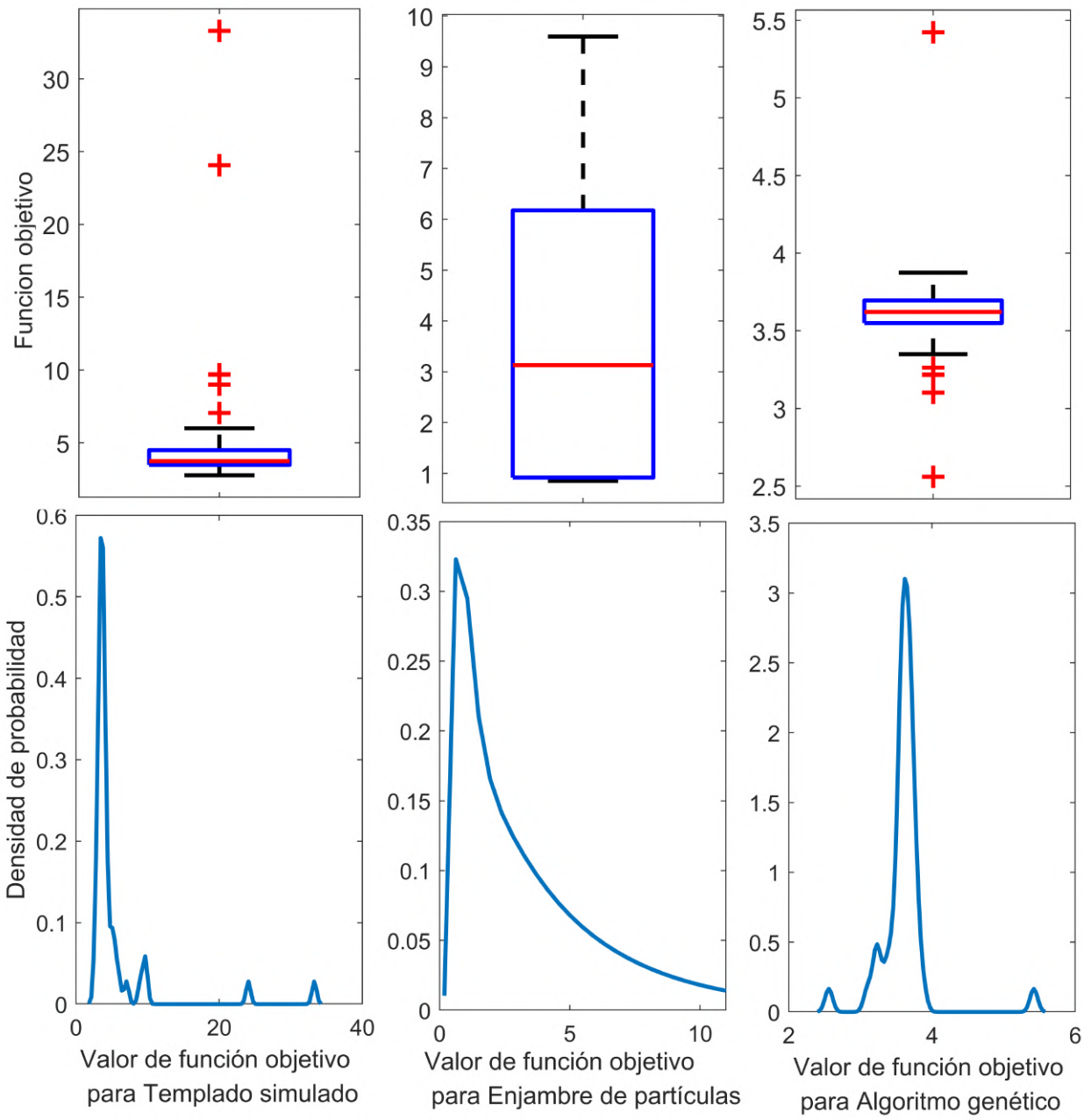


Figura 4-7: Desempeño de los optimizadores SA, PSO y GA con configuraciones por defecto y factor de normalización media.

se ubica alrededor de 0.68 que en comparación con un φ_{conf} igual a 0.018 (mejor valor encontrado con factor de normalización máximo en Tabla **4-5**) está lejos del óptimo global. También, GA presenta sus mejores resultados como puntos atípicos que no sobrepasan un valor φ_{conf} de 0.5. Por otra parte, los optimizadores PSO y SA presentan una distribución de φ_{conf} más amplia con una mediana alrededor de 0.6. Esto indica que el 50 % de los valores de función objetivo alcanzados por SA y PSO son mejores que los mejores resultados calculados

con GA. Sin embargo, la dispersión de φ_{conf} obtenida con el optimizador SA es casi la mitad del rango calculado con PSO, implicando una mejor precisión y reproducibilidad. Bajo factor de normalización máximo, el optimizador SA presenta la mejor relación exploración-explotación, seguido de PSO, mientras que GA muestra los peores resultados con inhabilidad de escapar de mínimos locales.

d) Configuraciones con normalización de valor medio

Los resultados de la dispersión de φ_{conf} y densidad de probabilidad para los algoritmos de optimización sintonizados con factor de normalización media son mostrados en la Figura **B-37**. El optimizador PSO presenta la dispersión de φ_{conf} más amplia de los tres optimizadores, con mediana alrededor de 1.0, en contraste con el optimizador SA que presentó una mediana alrededor de 1.1. EL optimizador GA muestra la dispersión de φ_{conf} más estrecha con mediana alrededor de 3.5 y mejor valor de 2.7 como punto atípico. Un fenómeno particular aparece para PSO, donde la mediana de φ_{conf} es mejor que el mejor valor generado en su proceso de sintonización, lo cual es causado por un mayor número de iteraciones permitidas. El algoritmo de optimización PSO presenta la mejor relación exploración-explotación bajo factor de normalización media con mayor precisión y reproducibilidad que el optimizador SA. GA presenta el peor resultado con inhabilidad para escapar de mínimos locales.

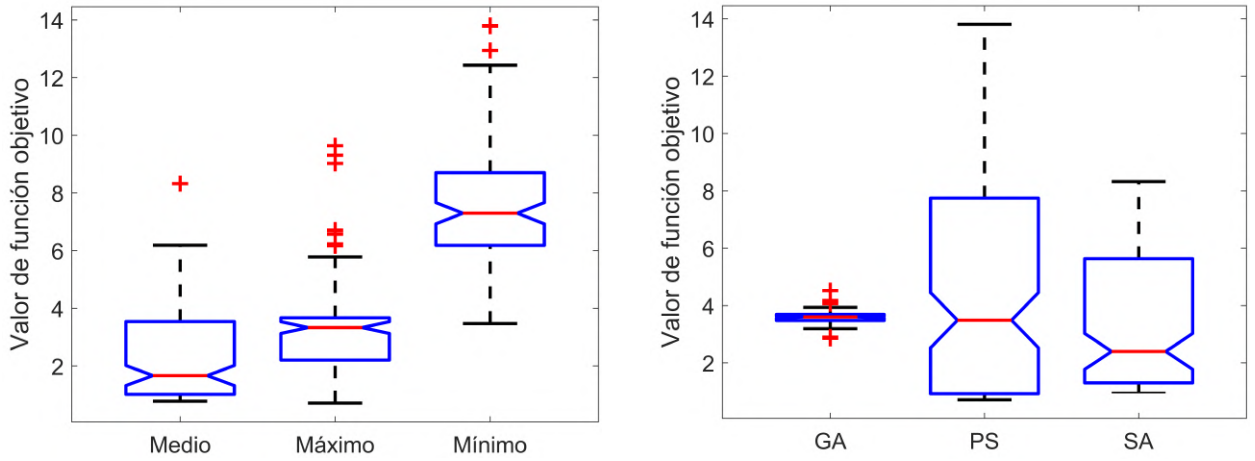
Pruebas estadísticas para comparación del desempeño para combinaciones de algoritmos de optimización sintonizados y factores de normalización

Los resultados estadísticos de la prueba no paramétrica de Friedman y la prueba de Dunn para la comparación del desempeño de las combinaciones de factores de normalización y algoritmos de optimización se muestra en la Tabla **4-9**. Existen diferencias significativas en el desempeño entre los factores de normalización y los optimizadores en la "Comparación A" y la "Comparación B". Esto significa que la combinación de algoritmo de optimización y factor de normalización tiene un fuerte efecto en el desempeño del optimizador en este problema de optimización específico. La prueba de Dunn afirma que los tres factores de normalización tienen efectos diferentes sobre el desempeño del optimizador con el factor de normalización mínimo presentando la mayor diferencia.

Con respecto a los algoritmos de optimización, el desempeño del optimizador GA es diferente tanto de PSO como de SA, los cuales presentan un desempeño similar. La Figura **4-8** muestra las distribuciones de φ_{conf} calculadas con factores de normalización (izquierda) y algoritmos de optimización (derecha). Estas distribuciones sirven como un complemento a las pruebas estadísticas de Friedman y Dunn. De la Figura **4-8**, el factor de normalización media presentan el mejor desempeño con todos los optimizadores, seguido del factor de normalización máximo. Además, los optimizadores PSO y SA presentan distribuciones estadísticamente similares para todos los factores de normalización.

Tabla 4-9: Pruebas de Friedman y Dunn para tipos de normalización de función objetivo con configuraciones sintonizadas para los algoritmos de optimización. GA - Algoritmo Genético, PSO - Enjambre de Partículas, SA - Templado Simulado.

Comparación A			Comparación B		
Friedman			Friedman		
Grupos		Normalización	Grupos		Optimizadores
Bloques		Optimizadores	Bloques		Normalización
Valor p		$1.6527\text{x}10^{-54}$	Valor p		$2.5177\text{x}10^{-9}$
Dunn			Dunn		
Comparaciones		Valor p	Comparaciones		Valor p
Máximo	Medio	$5.16\text{x}10^{-3}$	GA	PSO	0.0016
Máximo	Mínimo	$1.96\text{x}10^{-30}$	GA	SA	$5.93\text{x}10^{-5}$
Medio	Mínimo	$5.33\text{x}10^{-47}$	PSO	SA	0.178



(a) Tipo de normalización

(b) Optimizadores sintonizados

Figura 4-8: Dispersión de valores de función objetivo con optimizadores sintonizados y tipos de normalización. GA - Algoritmo Genético, PS - Enjambre de Partículas, SA - Templado Simulado.

Los resultados de las pruebas estadísticas confirman que el factor de normalización tiene papel relevante sobre el desempeño del algoritmo de optimización, inclusive cuando el mismo se encuentra sintonizado. Una posible explicación a este fenómeno es que el incremento en la escala de φ_{conf} genera un espacio de búsqueda más complejo debido a la aparición de más óptimos locales. Además, a pesar del uso lógico del factor de normalización máximo, se encontró que el factor de normalización media presenta resultados más robustos (Moles et al., 2003; Gahlawat & Srivastava, 2013). En general, el efecto del factor de normalización

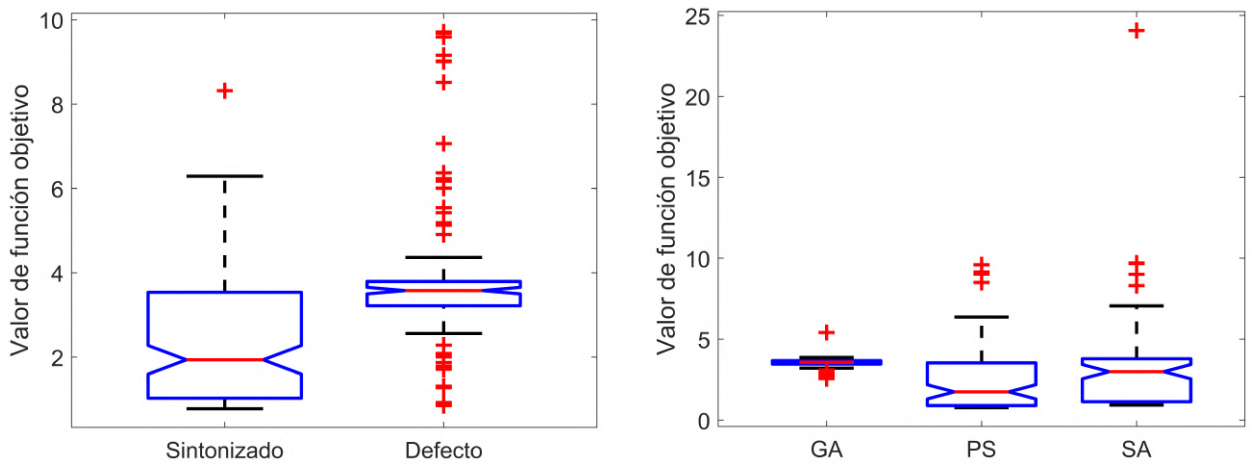
en el valor de la función objetivo ha sido poco estudiado, considerando que el factor de normalización es usado para reducir o eliminar el impacto de las diferencias en escala entre los datos experimentales usados como parte del problema de optimización (García et al., 2017).

Para comparar las diferencias en el desempeño de la optimización debido a la sintonización se realizó otra prueba estadística con un mismo factor de normalización. Los resultados de la comparación son mostrados en la Tabla **4-10**. Los resultados estadísticos muestran diferencias significativas entre las configuraciones sintonizadas y por defecto para los tres algoritmos de optimización (Comparación A). Además, existen diferencias significativas en el desempeño entre los tres optimizadores (Comparación B). La Figura **4-9** muestra la dispersión de φ_{conf} calculada para las configuraciones sintonizadas y por defecto (izquierda) y los optimizadores (derecha). Las configuraciones sintonizadas presentan los mejores resultados con mediana alrededor de 2.0 y por lo menos un 25 % de las optimizaciones realizadas tienen un valor de función objetivo inferior a 1.0. Con respecto a los algoritmos de optimización, PSO presenta el mejor desempeño con configuraciones sintonizadas y por defecto, seguido de SA y finalmente, GA presenta el peor desempeño.

Tabla 4-10: Pruebas de Friedman y Dunn para sintonización de algoritmos de optimización. GA - Algoritmo Genético, PS - Enjambre de Partículas, SA - Templado Simulado.

Comparación A			Comparación B		
Friedman			Friedman		
Grupos		Sintonización	Grupos		Optimizadores
Bloques		Optimizadores	Bloques		Sintonización
Valor p		$6.1944\text{x}10^{-11}$	Valor p		$5.6587\text{x}10^{-9}$
Dunn			Dunn		
Comparaciones		Valor p	Comparaciones		Valor p
			GA	PS	$5.34\text{x}10^{-9}$
Si	No	$3.9728\text{x}10^{-11}$	GA	SA	$2.37\text{x}10^{-3}$
			PS	SA	$0.09\text{x}10^{-3}$

La mejor combinación de factor de normalización y algoritmo de optimización, en este caso, corresponde a PSO sintonizado con factor de normalización media. Diferentes estudios han comparado el desempeño de metaheurísticas en estimación de parámetros de modelos de procesos biotecnológicos. Por ejemplo, un modelo dinámico no lineal de una red bioquímica ha sido investigado. El modelo cuenta con 8 ecuaciones diferenciales y 36 parámetros, donde se compararon 7 algoritmos de optimización con normalización de valor máximo (Moles et al., 2003). De estos optimizadores 4 corresponden al tipo de evolución diferencial (misma que GA). Los resultados indican que el mejor optimizador corresponde a estrategia evolutiva con clasificación estocástica el cual presento tanto el menor valor de función objetivo como esfuerzo computacional.



(a) Configuración (b) Algoritmos de optimización

Figura 4-9: Dispersión de valores de función objetivo de optimizadores con configuraciones sintonizadas y por defecto. GA - Algoritmo Genético, PS - Enjambre de Partículas, SA - Templado Simulado.

Roeva y Trenkova analizaron un modelo de fermentación de tipo fedbatch con *E. coli* para producción de ácido acético compuesto por 4 ecuaciones diferenciales y 8 parámetros, con dos metaheurísticas, GA y algoritmo de luciérnaga, sin normalización de función objetivo (Roeva & Trenkova, 2012). El algoritmo de luciérnaga pertenece a la categoría de inteligencia de enjambre (misma que PSO). Este estudio encontró un mejor desempeño del algoritmo de luciérnaga respecto a GA con menor valor de función objetivo y convergencia más veloz. Además, el desempeño de GA, algoritmo de luciérnaga y optimización por colonia de hormigas (inteligencia de enjambre) ha sido investigado para la estimación de parámetros de un modelo de fermentación fed-batch de *E. coli* con 3 estads y 5 parámetros sin factor de normalización (Roeva & Fidanova, 2014). Se encontró que la combinación del algoritmo de optimización por colonia de hormigas con algoritmo de luciérnaga o GA conserva la precisión de la estimación reduciendo significativamente el tiempo de cómputo.

Rocha *et al.* realizaron la estimación de parámetros o entradas de cuatro modelos matemáticos: producción de ácido acético (6 estados y 10 parámetros), producción de etanol (4 estados y 1 entrada a determinar), producción de anticuerpos monoclonales (7 estados y 16 parámetros) y producción de una proteína recombinante (7 estados y 2 entradas) (Rocha et al., 2014). El factor de normalización usado corresponde a punto experimental y se determinó que el algoritmo de evolución diferencial presenta mejores resultados en términos de valor de función objetivo y esfuerzo computacional que PSO. En general, la literatura disponible muestra un mejor desempeño de los algoritmos de tipo inteligencia de enjambre para estimación de parámetros de modelos matemáticos de procesos biotecnológicos. Sin embargo,

también se esperaría un buen desempeño de algoritmos de evolución diferencial como GA, lo cual no fue el caso en este trabajo.

Una explicación para el pobre desempeño del optimizador GA en el caso de estudio analizado y reflejado en su incapacidad para escapar de mínimos locales, es su falta de garantía de convergencia teórica en contraste con los optimizadores PSO y SA (Yang, 2000; Sun et al., 2012). Desde el punto de vista de un proceso de Markov, se ha demostrado que dependiendo de los operadores de mutación y selección utilizados y su combinación, el optimizador GA puede no converger al óptimo global ni siquiera en un tiempo infinito (Rudolph, 1994). Sin embargo, Thierens y Golberg analizaron las propiedades de convergencia del optimizador GA con diferentes operadores de selección y probaron que este algoritmo de optimización puede, de hecho, alcanzar convergencia teórica (Thierens & Goldberg, 1994). Schmitt revisó las propiedades de convergencia de diferentes combinaciones reportadas de operadores de selección y mutación (Schmitt, 2001). Para diferentes combinaciones de los operadores de mutación, selección y entrecruzamiento, no hay garantía de convergencia teórica, por lo menos para la mayoría de ellos. Esto demuestra que el optimizador GA tiene un problema crítico de convergencia teórica. Las combinaciones analizadas por Schmitt se presentan de forma detallada en la Tabla **B-4**.

En particular, operadores de selección basados en templado simulado y función de Boltzmann no están implementados en Matlab® R2018, los cuales son convergentes según el estudio de Schmitt. Tomando en cuenta las propiedades de convergencia exhibidas por el optimizador GA, se presenta una alta probabilidad de fallo para encontrar el óptimo global. Esto es una situación preocupante considerando el uso extenso de esta metaheurística en el área de procesos biotecnológicos. Ejemplos de lo anterior incluyen la estimación de parámetros de un modelo de *E. coli* (Roeva & Atanassova, 2016), entrenamiento de redes neuronales para optimización de producción de xilitol (Pappu & Gummadi, 2017), estimación de parámetros de un modelo de fermentación ABE (Díaz & Willis, 2018), estimación de parámetros de un modelo de fermentación etanólica (Prieto-Escobar et al., 2018), estimación de parámetros de un modelo de producción de biosurfactante a partir de cáscara de remolacha (Campos et al., 2018), uso de algoritmo genético en conjunto con filtros de Kalman para estimación de parámetros de un fotobiorreactor para producción de microalgas (García-Mañas et al., 2019) y estimación de parámetros de un modelo de producción de ramnolípidos biosurfactates (Câmara et al., 2020). Debe resaltarse que en ninguno de los trabajos mencionados se realizó sintonización de este optimizador antes de su uso.

En cuanto a la comparación de algoritmos de optimización realizada en este trabajo, es posible afirmar que fue realizada correctamente, con una cantidad aceptable de muestras por combinación de algoritmo y tipo de normalización junto con el uso de pruebas estadísticas no paramétricas (Chiarandini et al., 2007; LaTorre et al., 2020). Sin embargo, otras apro-

ximaciones para la sintonización de metaheurísticas han sido planteadas, ejemplo de esto incluyen una combinación de *racing* con superficie de respuesta para la sintonización de GA y SA, con una mejora significativa en el desempeño de estos optimizadores, especialmente SA (de Moraes Barbosa et al., 2015). Además, se tiene una aproximación con combinación de *racing* con diseño de experimentos aplicada a GA y SA, en donde SA nuevamente muestra un desempeño superior (de Moraes & Senne, 2017).

4.4.3. Directrices para sintonización de algoritmos de optimización

Los algoritmos de optimización global estocásticos, también llamados metaheurísticas, presentan una ventaja sobre sus partes determinísticas en cuanto a facilidad de implementación y velocidad de búsqueda. Sin embargo, al contrario de los optimizadores determinísticos, las metaheurísticas no presentan formulaciones matemáticas rigurosas que garanticen la precisión del resultado final (óptimo global). Una posible alternativa para mejorar la capacidad de búsqueda (exploración) y precisión (explotación) de las metaheurísticas corresponde a la modificación del valor de sus parámetros internos (o hiperparámetros). La sintonización de algoritmos de optimización estocásticos permite realizar un ajuste a los hiperparámetros del optimizador, lo que ingresa información del problema de optimización en el mecanismo de búsqueda haciéndolo más eficiente (relación exploración-explotación). Esta conclusión se deriva del teorema "no hay almuerzos gratis" (*no free lunch theorem* o NFLT por sus siglas en ingles) el cual establece dos condiciones: primero, un único mecanismo de búsqueda (algoritmo de optimización) tendrá diferente desempeño en diferentes problemas de optimización. Segundo, sino se incorpora información del problema de optimización en el optimizador, este no tendrá desempeño mejor que una búsqueda puramente aleatoria (Wolpert & Macready, 1997). En consideración del teorema NFLT y los resultados obtenidos en esta tesis, a continuación, se presentan las directrices para sintonización de algoritmos de optimización estocásticos basados en la Figura **4-1**.

1. Selección de algoritmo de optimización: La selección del algoritmo de optimización permite encontrar un mecanismo de búsqueda apropiado para el problema de optimización específico.

    1.1. Realizar como mínimo 13 optimizaciones con diferentes metaheurísticas en sus configuraciones por defecto.

    1.2. Aplicar una prueba no paramétrica de Kruskal-Wallis (si la única variable es el optimizador) o de Friedman (si se tienen otras variables a considerar).

    1.3. En caso de que los optimizadores presenten desempeños estadísticamente diferentes, se procede a realizar una prueba no paramétrica de Dunn.

    1.4. Seleccionar el optimizador con menor valor de mediana.

2. Sintonización de hiperparámetros: el proceso de sintonización permite incorporar información del problema de optimización en el mecanismo de búsqueda, lo que aumenta la eficiencia, reproducibilidad y precisión de la metaheurística (relación exploración-explotación). Este procedimiento se realiza mediante la herramienta híbrida para sintonización de metaheurísticas por interconexión de Matlab® - R® (Figura **4-2**). Los paquetes "irace" y "R.matlab" deben estar preinstalados en R®.

   2.1. Especificar los rangos de los hiperparámetros e incluirlos en el archivo de texto `parameters.txt`.

   2.2. Especificar las opciones del algoritmo "irace" (cantidad máxima de pruebas, estadísticos de descarte, entre otros) en el archivo de texto `scenario.txt`.

   2.3. Especificar el problema de optimización en el archivo de Matlab® `Optimizer.m`.

   2.4. Ejecutar el algoritmo "irace".

   2.5. Seleccionar la mejor configuración élite obtenida.

4.5. Conclusiones

En este trabajo, se implementó satisfactoriamente una herramienta híbrida para la sintonización de algoritmos de optimización usando R® y Matlab® a través de los paquetes de R® "R.matlab" (comunicación entre R® y Matlab®) y "irace" (algoritmo de sintonización), además de proveer una comparación sistemática del desempeño de la sintonización. Particularmente, fue investigada la influencia del factor de normalización de la función objetivo y la sintonización de los parámetros internos del algoritmo de optimización en la habilidad del optimizador para explorar el espacio de búsqueda (relación exploración-explotación). Una comparación exhaustiva entre tres algoritmos de optimización (PSO, GA y SA) y tres factores de normalización (media, mínimo y máximo de estado experimental) fue realizada.

Los resultados de esta investigación muestran que la sintonización de algoritmos de optimización debe ser un paso fundamental siempre que se requiera del uso de metaheurísticas que posean parámetros internos. Como se demostró en este capítulo, la configuración de un algoritmo de optimización estocástico tiene un papel fundamental en el desempeño del optimizador y la reproducibilidad y precisión de los resultados obtenidos.

Para el caso de estudio de bioproducción de xilitol se encontró que el factor de normalización más eficiente corresponde a la media del estado experimental por su capacidad de ponderar de manera consistente en todos los estados involucrados en la función objetivo. Adicionalmente, la mejor combinación de factor de normalización y algoritmo de optimización encontrada fue PSO y factor de normalización media, con alta precisión y reproducibilidad en el valor de

la función objetivo. Por otra parte, el optimizador GA presentó el peor desempeño considerando su incapacidad para escapar de mínimos locales, lo cual es atribuido a su falta de garantía de convergencia teórica. Finalmente, las directrices para sintonización de algoritmos de optimización utilizados en modelos de procesos biotecnológicos se muestran en la Figura **4-1**.

Bibliografía

Aguiar e Oliveira, H., Ingber, L., Petraglia, A., Petraglia, M. R., & Machado, M. A. S. (2012). Stochastic global optimization and its applications with fuzzy adaptive simulated annealing. Springer Publishing Company, Incorporated.

Banga, J. R., Moles, C. G., & Alonso, A. A. (2004). Global optimization of bioprocesses using stochastic and hybrid methods. In Frontiers in global optimization (pp. 45–70). Springer.

Bartz-Beielstein, T. & Zaefferer, M. (2017). Model-based methods for continuous and discrete global optimization. Applied Soft Computing, 55, 154–167.

Ben-Eli, M. U. (2018). Sustainability: definition and five core principles, a systems perspective. Sustainability Science, 13(5), 1337–1343.

Câmara, J. M., Sousa, M. A., & Barros Neto, E. L. (2020). Modeling of rhamnolipid biosurfactant production: Estimation of kinetic parameters by genetic algorithm. Journal of Surfactants and Detergents, 23(4), 705–714.

Cameron, I. T. & Hangos, K., Eds. (2001). Process Modelling and Model Analysis, volume 4. Academic press.

Campos, A., Nogueira, J., Coelho, F., Fileti, A., & Santos, B. (2018). Genetic algorithm optimization of the parameters involved in biosurfactant production from beet peel as substrate. Chemical Engineering Transactions, 65, 469–474.

Charaniya, S., Hu, W.-S., & Karypis, G. (2008). Mining bioprocess data: opportunities and challenges. Trends in biotechnology, 26(12), 690–699.

Chen, J., Xin, B., Peng, Z., Dou, L., & Zhang, J. (2009). Optimal contraction theorem for exploration–exploitation tradeoff in search and optimization. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 39(3), 680–691.

Chiarandini, M., Paquete, L., Preuss, M., & Ridge, E. (2007). Experiments on metaheuristics: Methodological overview and open issues. Technical report, University of Southern Denmark.

Couceiro, M. & Ghamis, P. (2016). Fractional Order DarwinianParticle Swarm Optimization: Applications and Evaluationof an Evolutionary Algorithm. Springer.

Coy, S. P., Golden, B. L., Runger, G. C., & Wasil, E. A. (2001). Using experimental design to find effective parameter settings for heuristics. Journal of Heuristics, 7(1), 77–97.

de Moraes, Barbosa, E. B. & Senne, E. L. F. (2017). Improving the fine-tuning of metaheuristics: an approach combining design of experiments and racing algorithms. Journal of Optimization, (pp. 1 – 7).

de Moraes Barbosa, E. B., Senne, E. L. F., & Silva, M. B. (2015). Improving the performance of metaheuristics: an approach combining response surface methodology and racing algorithms. International journal of engineering mathematics, (pp. 1–9).

Desing, H., Braun, G., & Hischier, R. (2020). Ecological resource availability: a method to estimate resource budgets for a sustainable economy. Global Sustainability, 3.

Díaz, V. H. G. & Willis, M. J. (2018). Kinetic modelling and simulation of batch, continuous and cell-recycling fermentations for acetone-butanol-ethanol production using clostridium saccharoperbutylacetonicum n1-4. Biochemical Engineering Journal, 137, 30–39.

Dinno, A. (2017). Package "dunn.test", Dunn's Test of Multiple Comparisons Using Rank Sums.

Dowsland, K. A. & Thompson, J. (2012). Handbook of natural computing, chapter Simulated annealing, (pp. 1623–1655). Springer-Verlag.

Ezugwu, A. E., Shukla, A. K., Nath, R., Akinyelu, A. A., Agushaka, J. O., Chiroma, H., & Muhuri, P. K. (2021). Metaheuristics: a comprehensive overview and classification along with bibliometric analysis. Artificial Intelligence Review, (pp. 1–80).

Fahoome, G. & Sawilowsky, S. S. (2002). Review of twenty nonparametric statistics and their large sample approximations. Journal of Modern Applied Statistical Methods, 1(2), 248–268.

Gahlawat, G. & Srivastava, A. K. (2013). Development of a mathematical model for the growth associated polyhydroxybutyrate fermentation by azohydromonas australica and its use for the design of fed-batch cultivation strategies. Bioresource technology, 137, 98–105.

García, M. R., Alonso, A. A., & Balsa-Canto, E. (2017). A normalisation strategy to optimally design experiments in computational biology. In International Conference on Practical Applications of Computational Biology & Bioinformatics (pp. 126–136).: Springer.

García-Mañas, F., Guzmán, J., Berenguel, M., & Acién, F. (2019). Biomass estimation of an industrial raceway photobioreactor using an extended kalman filter and a dynamic model for microalgae production. Algal Research, 37, 103–114.

Huang, C., Li, Y., & Yao, X. (2019). A survey of automatic parameter tuning methods for metaheuristics. IEEE Transactions on Evolutionary Computation, 24(2), 201–216.

Hutter, F., Hoos, H. H., & Leyton-Brown, K. (2011). Sequential model-based optimization for general algorithm configuration. In International conference on learning and intelligent optimization (pp. 507–523).: Springer.

Jasrasaria, D. & Pyzer-Knapp, E. O. (2018). Dynamic control of explore/exploit trade-off in bayesian optimization. In Science and Information Conference (pp. 1–15).: Springer.

Ji, Y. (2012). Model based process design for bioprocess optimisation: case studies on precipitation with its applications in antibody purification. PhD thesis, UCL (University College London).

Koutinas, M., Kiparissides, A., Pistikopoulos, E. N., & Mantalaris, A. (2012). Bioprocess systems engineering: transferring traditional process engineering principles to industrial biotechnology. Computational and structural biotechnology journal, 3(4), e201210022.

Kramer, O. (2017). Genetic algorithm essentials, volume 679. Springer.

LaTorre, A., Molina, D., Osaba, E., Del Ser, J., & Herrera, F. (2020). Fairness in bio-inspired optimization research: A prescription of methodological guidelines for comparing meta-heuristics. arXiv preprint to Swarm and Evolutionary Computation.

Lecchini-Visintini, A., Lygeros, J., & Maciejowski, J. M. (2010). Stochastic optimization on continuous domains with finite-time guarantees by markov chain monte carlo methods. IEEE Transactions on Automatic Control, 55(12), 2858–2863.

L'Ecuyer, P. (2012). Random number generation. In Handbook of computational statistics (pp. 35–71). Springer.

Lim, S. M., Sultan, A. B. M., Sulaiman, M. N., Mustapha, A., & Leong, K. Y. (2017). Crossover and mutation operators of genetic algorithms. International journal of machine learning and computing, 7(1), 9–12.

López-Ibáñez, M., Dubois-Lacoste, J., Cáceres, L. P., Birattari, M., & Stützle, T. (2016). The irace package: Iterated racing for automatic algorithm configuration. Operations Research Perspectives, 3, 43–58.

Mirjalili, S., Dong, J. S., & Lewis, A. (2020). Nature-inspired optimizers: theories, literature reviews and applications, volume 811 of Studies in Computational Intelligence. Springer, 1 edition.

Moles, C. G., Mendes, P., & Banga, J. R. (2003). Parameter estimation in biochemical pathways: a comparison of global optimization methods. Genome research, 13(11), 2467–2474.

Montana-Hoyos, C. & Fiorentino, C. (2016). Bio-utilization, bio-inspiration, and bio-affiliation in design for sustainability: Biotechnology, biomimicry, and biophilic design. The International Journal of Designed Objects, 10(3), 1–18.

Pappu, S. M. J. & Gummadi, S. N. (2017). Artificial neural network and regression coupled genetic algorithm to optimize parameters for enhanced xylitol production by debaryomyces nepalensis in bioreactor. Biochemical Engineering Journal, 120, 136–145.

Poli, R., Kennedy, J., & Blackwell, T. (2007). Particle swarm optimization. Swarm intelligence, 1(1), 33–57.

Prieto-Escobar, N., Saldarriaga-Aristizábal, P. A., & Chaparro-Muñoz, V. (2018). Heuristic parameter estimation for a continuous fermentation bioprocess. Revista Facultad de Ingeniería Universidad de Antioquia, (88), 26–39.

Quinn, G. P. & Keough, M. J. (2002). Experimental design and data analysis for biologists. Cambridge university press.

Riff, M.-C. & Montero, E. (2013). A new algorithm for reducing metaheuristic design effort. In 2013 IEEE Congress on Evolutionary Computation (pp. 3283–3290).: IEEE.

Rocha, M., Mendes, R., Rocha, O., Rocha, I., & Ferreira, E. C. (2014). Optimization of fed-batch fermentation processes with bio-inspired algorithms. Expert Systems with Applications, 41(5), 2186–2195.

Roeva, O. & Atanassova, V. (2016). Cuckoo search algorithm for model parameter identification. International Journal Bioautomation, 20(4), 483 – 492.

Roeva, O. & Fidanova, S. (2014). Parameter identification of an e. coli cultivation process model using hybrid metaheuristics. International Journal of Metaheuristics 9, 3(2), 133–148.

Roeva, O. & Trenkova, T. V. (2012). Genetic algorithms and firefly algorithms for non-linear bioprocess model parameters identification. In IJCCI (pp. 164–169).

Rudolph, G. (1994). Convergence analysis of canonical genetic algorithms. IEEE transactions on neural networks, 5(1), 96–101.

Sakawa, M. (2012). Genetic algorithms and fuzzy multiobjective optimization, volume 14. Springer Science & Business Media.

Schmitt, L. M. (2001). Theory of genetic algorithms. Theoretical Computer Science, 259(1-2), 1–61.

Siddique, N. & Adeli, H. (2016). Simulated annealing, its variants and engineering applications. International Journal on Artificial Intelligence Tools, 25(06), 1630001.

Sirisansaneeyakul, S., Wannawilai, S., & Chisti, Y. (2013). Repeated fed-batch production of xylitol by candida magnoliae tistr 5663. Journal of Chemical Technology & Biotechnology, 88(6), 1121–1129.

Spall, J. C. (2012). Stochastic optimization. In Handbook of computational statistics (pp. 173–201). Springer.

Stephanopoulos, G. & Reklaitis, G. V. (2011). Process systems engineering: From solvay to modern bio-and nanotechnology.: A history of development, successes and prospects for the future. Chemical engineering science, 66(19), 4272–4306.

Sun, J., Wu, X., Palade, V., Fang, W., Lai, C.-H., & Xu, W. (2012). Convergence analysis and improvements of quantum-behaved particle swarm optimization. Information Sciences, 193, 81–103.

Thierens, D. & Goldberg, D. (1994). Convergence models of genetic algorithm selection schemes. In International Conference on Parallel Problem Solving from Nature (pp. 119–129).: Springer.

Tochampa, W., Sirisansaneeyakul, S., Vanichsriratana, W., Srinophakun, P., Bakker, H. H., Wannawilai, S., & Chisti, Y. (2015). Optimal control of feeding in fed-batch production of xylitol. Industrial & Engineering Chemistry Research, 54(7), 1992–2000.

Ur-Rehman, S., Mushtaq, Z., Zahoor, T., Jamil, A., & Murtaza, M. A. (2015). Xylitol: a review on bioproduction, application, health benefits, and related safety issues. Critical reviews in food science and nutrition, 55(11), 1514–1528.

Villaverde, A. F., Fröhlich, F., Weindl, D., Hasenauer, J., & Banga, J. R. (2019). Benchmarking optimization methods for parameter estimation in large kinetic models. Bioinformatics, 35(5), 830–838.

Wang, D., Tan, D., & Liu, L. (2018a). Particle swarm optimization algorithm: an overview. Soft Computing, 22(2), 387–408.

Wang, N., Phelan, P. E., Harris, C., Langevin, J., Nelson, B., & Sawyer, K. (2018b). Past visions, current trends, and future context: A review of building energy, carbon, and sustainability. Renewable and Sustainable Energy Reviews, 82, 976–993.

Wolpert, D. H. & Macready, W. G. (1997). No free lunch theorems for optimization. IEEE transactions on evolutionary computation, 1(1), 67–82.

Yang, R. (2000). Convergence of the simulated annealing algorithm for continuous global optimization. Journal of optimization theory and applications, 104(3), 691–716.

Zhigljavsky, A. & Zilinskas, A. (2007). Stochastic global optimization, volume 9. Springer Science & Business Media.

Capítulo 5

Estimación práctica de parámetros

5.1. Resumen

La estimación numérica o *identificabilidad práctica* de parámetros es un proceso que involucra diversos elementos: modelo matemático, función objetivo, optimizador, solucionador y datos experimentales, que unidos permiten *estimar* el valor de los parámetros desconocidos de un modelo matemático, dotando a este de capacidad descriptiva sobre el sistema modelado. En el presente capítulo se investiga el efecto de la interconexión entre los diferentes componentes del problema de estimación de parámetros en el valor calculado de los parámetros. Así mismo, se analizará la capacidad tanto descriptiva como predictiva y la interpretabilidad de los parámetros estimados usando como caso de estudio el modelo matemático de fermentación diaúxica de glucosa y xilosa para producción de xilitol. Como resultado se propone entonces una metodología general o *framework* para la estimación de parámetros de modelos matemáticos de procesos biotecnológicos diseñada para hacer viable la consecución de modelos matemáticos con interpretabilidad en sus parámetros y estadísticamente confiables, los cuales son particularmente importantes para el diseño, control y optimización de procesos bajo la filosofía de ingeniería de sistemas de proceso (PSE). La contribución "Parameter estimation and model validation in bioprocess mathematical models: an extensive methodology" que presenta los resultados de este capítulo se encuentra en construcción.

5.2. Introducción

El continuo crecimiento de la población humana a nivel mundial ha conllevado a una demanda creciente de recursos naturales y energía, a la par de un aumento en la degradación ambiental por la generación de contaminantes como los gases de efecto invernadero (Khan et al., 2020). La biotecnología puede ofrecer una solución a estas problemáticas mediante la generación de procesos industriales más eficientes y ambientalmente amigables basados en el uso de organismos vivos o partes de ellos (Katz et al., 2018). Desde la perspectiva de PSE

es necesario describir los sistemas biológicos en forma de modelos matemáticos útiles para el diseño, control y optimización de procesos.

Los modelos matemáticos de procesos biotecnológicos son en general modelos de tipo "caja gris", lo que implica una combinación entre conocimientos teóricos acerca de las relaciones entre estados y términos cuyo valor no se conoce *a priori* llamados parámetros (González-Figueredo et al., 2018). Diferentes combinaciones de valores para los parámetros conducen a soluciones y tendencias particulares del modelo matemático. Los parámetros son entonces elementos dentro del modelo que pueden capturar las particularidades del sistema analizado, con el fin de que el modelo tenga tanto capacidad descriptiva como predictiva (Koutinas et al., 2012). El proceso de calcular los valores de los parámetros se conoce como ***identificabilidad práctica***, en la cual se involucran el modelo matemático y datos experimentales con el objetivo de aproximar, en la medida de lo posible, el comportamiento del modelo al comportamiento del sistema (Dochain, 2013). Lo anterior puede lograrse a través de la solución de un *problema de optimización*, el cual se compone de los siguientes elementos: modelo matemático, función objetivo, datos experimentales, algoritmo de optimización, solucionador, variables de optimización y restricciones (Aster et al., 2005; Biegler, 2010).

De forma general, se requiere la solución de problemas de optimización no lineales y no convexos, como es el caso más común en modelos matemáticos de procesos biotecnológicos. Dichos modelos presentan relaciones no lineales entre los parámetros del modelo, tales como funciones racionales (e.g. cinéticas de tipo Michaelis-Menten y Monod) y multilinealidades (multiplicación de parámetros), entre otros (Englezos & Kalogerakis, 2000; DiStefano III, 2015). Además, funciones objetivo no convexas implican la presencia de múltiples soluciones (máximos o mínimos) locales y una (posible) solución global. Lo anterior presenta un reto desde el punto de vista matemático, pues se requiere hallar la solución global que contendría el "valor verdadero" de los parámetros del modelo matemático analizado. Una vez estimado el valor de los parámetros se requiere un paso adicional, *validar* la calidad del modelo. La validación corresponde al proceso de cuantificar la *incertidumbre* presente en los parámetros estimados y como esta afecta a la predicción del modelo (Kroll et al., 2017; Sadino-Riquelme et al., 2020).

En los capítulos anteriores se definieron los elementos que componen el problema de optimización: modelo matemático en el capítulo 1, datos experimentales en el capítulo 2, identificabilidad estructural en el capítulo 3 y finalmente, los efectos de la función objetivo y algoritmo de optimización en el capítulo 4. Considerando lo anterior, se ha determinado que los parámetros del modelo matemático de fermentación diaúxica de glucosa y xilosa para producción de xilitol pueden ser conocidos, el optimizador más eficiente es el optimizador enjambre de partículas y su sintonización mejora de forma significativa la reproducibilidad de los resultados, se eliminaron puntos atípicos y ruido aleatorio en los datos experimentales

y la mejor normalización de función objetivo corresponde al valor medio de las variables experimentales.

El presente capítulo tiene por objetivo la consolidación de un marco metodológico o "framework" que permita la sintonización de modelos matemáticos de procesos biotecnológicos con capacidad descriptiva y predictiva, usando como referencia el modelo matemático de caso de estudio. Previamente, Cameron y Hangos propusieron una metodología enfocada en la construcción y estimación de parámetros de modelos matemáticos dinámicos (Cameron & Hangos, 2001). Sin embargo, en este trabajo se propone una metodología extendida para abarcar características propias de modelos matemáticos de procesos biotecnológicos como alta no linealidad, carencia de identificabilidad estructural y elevada incertidumbre en medidas experimentales. La metodología desarrollada considera la limpieza de datos experimentales, análisis de identificabilidad estructural, sintonización de algoritmos de optimización y análisis de incertidumbre en parámetros y respuesta del modelo matemático. Un modelo matemático que ha sido calibrado y validado se convierte, por lo tanto, en una herramienta sumamente importante para el entendimiento y aplicación de sistemas biológicos en PSE.

5.3. Metodología

A continuación, se presenta el marco metodológico o "framework" propuesto en esta investigación para estimación de parámetros y validación de modelos matemáticos de procesos biotecnológicos (Figura **5-1**), el cual fue aplicado al caso de estudio de un modelo matemático de fermentación diaúxica para bioproducción de xilitol. La metodología descrita se basa en procesos de limpieza de datos, análisis de identificabilidad estructural, sintonización de algoritmos de optimización, identificabilidad práctica y validación de modelos, los cuales requieren como insumo datos experimentales, modelo matemático, espacio paramétrico, función objetivo, solucionador, algoritmo de optimización, indicadores de ajuste, intervalos de confianza en parámetros y predictor e indices de sensibilidad. Este marco metodológico corresponde a una extensión de otros framework desarrollados previamente como el de Cameron & Hangos (2001) o Ljung (1990), el cual busca mejorar la interpretabilidad de los parámetros y la calidad del modelo, al considerar el efecto del ruido aleatorio en los datos experimentales, la identificabilidad estructural del modelo matemático, la precisión y reproducibilidad del optimizador y la calidad tanto descriptiva como predictiva del modelo matemático.

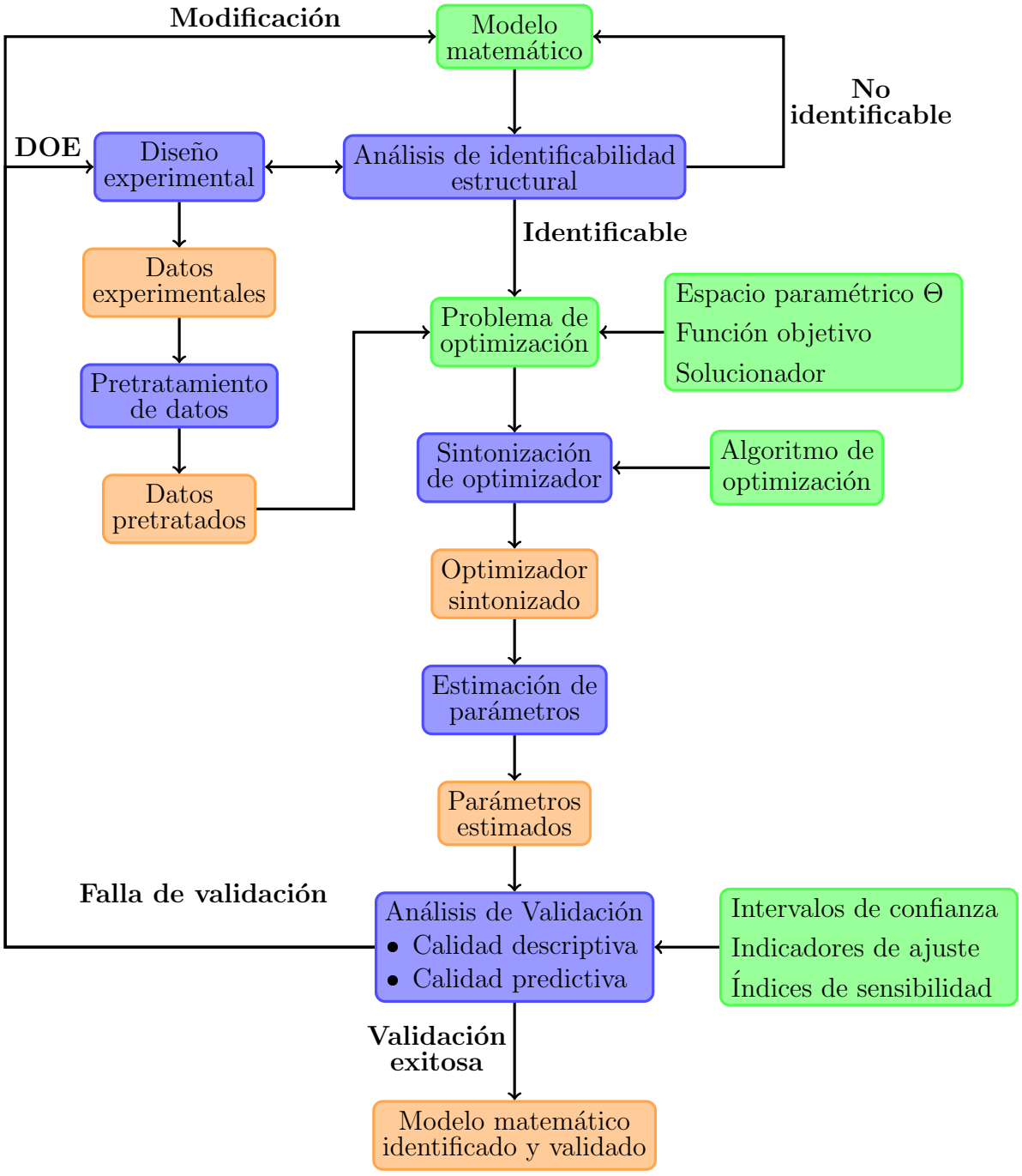


Figura 5-1: Metodología de estimación y validación de parámetros para modelos matemáticos de procesos biotecnológicos. ● insumo, ● proceso, ● resultado. DOE: diseño óptimo de experimentos.

5.3.1. Problema de optimización

La estimación de parámetros de procesos biotecnologícos puede plantearse como el siguiente problema de optimización:

$$min\ \varphi(\boldsymbol{\theta}) \tag{5-1}$$

$$s.t.\ \frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta}), \qquad \mathbf{x}_0 = \mathbf{x}(t_0) \tag{5-2}$$

$$\mathbf{g}(\mathbf{x}(t), \boldsymbol{\theta}) = 0 \tag{5-3}$$

$$\mathbf{h}(\mathbf{x}(t)) \leq 0 \tag{5-4}$$

$$\boldsymbol{\theta}_L \leq \boldsymbol{\theta} \leq \boldsymbol{\theta}_U, \tag{5-5}$$

$$\mathbf{x}_L \leq \mathbf{x} \leq \mathbf{x}_U \tag{5-6}$$

en donde $\varphi(\boldsymbol{\theta})$ corresponde a la función a minimizar (función objetivo), $\boldsymbol{\theta}$ vector de parámetros del modelo matemático del proceso biotecnológico, $\mathbf{f}$ vector de ecuaciones diferenciales, $\mathbf{g}$ vector de ecuaciones algebraicas, $\mathbf{h}$ vector de restricciones, $\mathbf{x}$ vector de estados, t tiempo y $\mathbf{x_0}$ vector de condiciones iniciales para ecuaciones diferenciales (Biegler, 2010). Los subindices L y U describen los valores de las cotas inferior y superior, respectivamente.

5.3.2. Modelo matemático

El caso de estudio seleccionado corresponde al modelo matemático de fermentación diaúxica de glucosa y xilosa para producción de xilitol en reactor batch propuesto por Tochampa et al. (2015). El modelo se resume en las Ecuaciones 5-7 a 5-14. El modelo es explicado de manera extensa en el capítulo 1, sección 1.4

$$\frac{dC_X}{dt} = \mu C_X \tag{5-7}$$

$$\frac{dC_{glu}}{dt} = -\left[q_{glu}^{max} \frac{C_{glu}}{C_{glu} + K_{S,glu}\left(1 + \dfrac{C_{xil}}{K_{i,xil}}\right)} \right] C_X \tag{5-8}$$

$$\frac{dC_{xil}}{dt} = -\left[q_{xil}^{max} \frac{C_{xil}}{C_{xil} + K_{S,xil}\left(1 + \dfrac{C_{glu}}{K_{i,glu}}\right)} \right] C_X \tag{5-9}$$

$$\frac{dC_{xit}^{in}}{dt} = \rho_X(r_{f,xit} - r_{u,xit} - r_{t,xit}) - \mu C_{xit}^{in} \tag{5-10}$$

$$\frac{dC_{xit}^{ex}}{dt} = r_{t,xit} C_X \tag{5-11}$$

$$\tag{5-12}$$

$$\mu = \mu_{glu}^{max} \frac{C_{glu}}{K_{S,glu} + C_{glu}} + \mu_{xit}^{max} \frac{C_{xit}^{in}}{K_{S,xit} + C_{xit}^{in}} \frac{K_r}{K_r + C_{glu}} \tag{5-13}$$

$$r_{t,xit} = 3.6\text{x}10^6 P_{xit} a_{cell} (C_{xit}^{in} - C_{xit}^{ex}) \tag{5-14}$$

5.3.3. Solucionadores

Dado que el modelo de fermentación diaúxica para producción de xilitol corresponde a un sistema ODE se requiere del uso de un integrador (una descripción más detallada de los integradores se da en el apéndice C.1). Debido a que los sistemas ODE pueden describir de manera simultánea fenómenos que ocurren en diferentes escalas de tiempo, estos sistemas se pueden comportar de manera rígida, caso común en modelos de procesos biotecnológicos. En el estudio realizado por Postawa *et al.* se probaron diferentes algoritmos de integración en un modelo matemático de producción de biogas (de tipo ODE), en donde el algoritmo `ode15s()` de Matlab demostró alta eficiencia, precisión y consistencia en los resultados (Postawa et al., 2020). Este algoritmo propuesto por Shampine & Reichelt (1997) es un integrador de tipo implícito, de tamaño de paso variable y A-estable que es robusto ante sistemas rígidos. Dado lo anterior, este fue el algoritmo de integración utilizado en esta investigación.

5.3.4. Función objetivo

La función objetivo es una $\varphi(\boldsymbol{\theta})$ es una medida de la desviación general de la salida calculada por el modelo matemático respecto a los datos experimentales. Mayor información acerca de funciones objetivo se muestra en el apéndice C.2. Dado que los parámetros del modelo matemático son desconocidos, es deseable que la función objetivo lleve a un estimado que se aproxime de la manera más probable al valor verdadero de los parámetros, es decir, $\varphi(\boldsymbol{\theta})$ debe corresponder a una función de máxima verosimilitud (Ljung, 1990). Comúnmente se recurre a la función de mínimos cuadrados, la cual posee las propiedades deseables de ser una función de máxima verosimilitud y no requerir de la suposición o cálculo de parámetros internos como los pesos w_i de la función QUAL2Kw o el factor α de una función de optimización regularizada (Aster et al., 2005; Pelletier et al., 2006). Adicionalmente, la información estadística de los datos experimentales puede introducirse de manera simple en caso de estar disponible.

Sí existen diferencias en la escala en los datos experimentales, estas pueden ser corregidas en la función de mínimos cuadrados como se analizó en el capítulo 4 de esta tesis. Debido a lo anterior, la función objetivo utilizada para la estimación de los parámetros del modelo matemático para bioproducción de xilitol es mínimos cuadrados con normalización por valor medio de los estados (Fórmula 5-15), en donde $\mathbf{y}(\mathbf{t})_i$ corresponde al dato experimental, $\mathbf{e}$ error de estimación o también llamado residual, N_{exp} número de datos experimentales por variable experimental, N_{var} número de variables experimentales y $\overline{\omega}$ valor medio de la variable experimental. Adicionalmente, se consideran conjuntos experimentales con diferente

cantidad de datos en donde N_{set} es el número de conjuntos de datos utilizados en la estimación.

$$\varphi(\boldsymbol{\theta}) = \sum_{k=1}^{N_{set}} \frac{1}{N_{var} \cdot N_{exp}} \sum_{j=1}^{N_{var}} \sum_{i=1}^{N_{exp}} \frac{\mathbf{e}_{i,j,k}^2}{\overline{\omega}_{j,k}} \qquad \mathbf{e}_i = \mathbf{x}(\boldsymbol{\theta}, t_i) - \mathbf{y}_i \tag{5-15}$$

5.3.5. Datos experimentales

La información experimental utilizada tanto para estimación de parámetros como validación del modelo corresponde a 22 experimentos consecutivos de tipo batch, tomados del trabajo de Sirisansaneeyakul et al. (2013). De manera específica, el experimento 1 cuenta con concentraciones iniciales tanto de glucosa como xilosa, en contraste, los experimentos 2 a 22 poseen únicamente xilosa como sustrato.

- Estimación de parámetros:
  Dada la naturaleza consecutiva de los experimentos, los conjuntos de datos 1 a 9 fueron utilizados para estimación de parámetros. Adicionalmente, para determinar el efecto de la eliminación de puntos atípicos y ruido a aleatorio en el valor de los parámetros estimados, se realizaron 45 optimizaciones con el conjunto de datos 1, tanto con el pretratamiento descrito en el capítulo 2 como sin pretratamiento.

- Validación:
  Para determinar la capacidad predictiva del modelo matemático del caso de estudio con parámetros estimados, es necesario utilizar información experimental adicional a la empleada en la estimación de parámetros. Por este motivo, los análisis de incertidumbre fueron realizados con los conjuntos de datos 10 a 22.

5.3.6. Algoritmo de optimización global

En el capítulo 4 se determinó que el mejor algoritmo de optimización global en el caso de estudio de bioproducción de xilitol corresponde a enjambre de partículas, presente en el *Global optimization toolbox* de Matlab® R2018 (Matlab, 2018). La configuración sintonizada de este optimizador se muestran en la Tabla **5-1**.

Tabla 5-1: Parámetros sintonizados del optimizador enjambre de partículas.

Parámetro	Valor	Parámetro	Valor
InertiaRangelb	0.2517	SelfAdjusWeight	1.5869
InertiaRangeub	0.5043	SocialAdjusmentweight	1.4422
MinNeighborsFraction	0.3559	SwarmSize	185

5.3.7. Delimitación del espacio paramétrico

La búsqueda de parámetros debe ser realizada en un espacio paramétrico Θ lo suficientemente amplio para asegurar que este contenga la combinación deseada de parámetros. Sin embargo, la definición exacta de los límites de Θ sigue siendo un problema abierto. De forma general, Θ se define con conocimiento *a priori* de los parámetros, en donde la bibliografía puede ofrecer indicios de valores posibles pero cuyos límites se verán afectados por múltiples variables como especie de microorganismo, medio de cultivo, tipo de fermentación, entre otros, que en la práctica adicionan incertidumbre a la cota por establecer (Tochampa et al., 2015; Mohamad et al., 2016; Dorantes-Landa et al., 2020). Debido a esto, habitualmente se opta por elegir un intervalo amplio a riesgo de incluir una mayor cantidad de mínimos locales y efectos de interacción entre los parámetros aunque otras alternativas han sido propuestas (Dayarian et al., 2009; Pitt & Banga, 2019). El espacio Θ definido para este trabajo es mostrado en la Tabla **5-2**. Estos límites fueron construidos con base en los valores estimados de los parámetros cuya combinación alcanzó un valor de función objetivo inferior a 0.9261, correspondiente a la mediana mostrada en la Figura **B-37** para el optimizador enjambre de partículas.

Tabla 5-2: Espacio paramétrico Θ.

Parámetro	Límite inferior	Límite superior	Parámetro	Límite inferior	Límite superior
μ_{glu}^{max}	1x10^{-3}	1x10^{-1}	q_{glu}^{max}	1x10^{-1}	100
$K_{S,\,glu}$	1x10^{-7}	100	q_{xil}^{max}	1x10^{-2}	100
μ_{xit}^{max}	1x10^{-2}	1.0	$K_{i,\,glu}$	1x10^{-9}	1.0
$K_{S,\,xil}$	1x10^{-5}	100	$K_{i,\,xil}$	1x10^{-6}	100
$K_{S,\,xit}$	1x10^{-8}	1.0	P_{xit}	1x10^{-9}	1x10^{-8}
K_r	1x10^{-5}	100			

De forma particular, puede presentarse la situación de que el valor estimado de un parámetro se encuentre sobre la cota impuesta (i.e. se tiene una restricción activa). Lo anterior puede atribuirse a dos situaciones: el espacio paramétrico no es lo suficientemente amplio, o bien, la interacción entre parámetros genera combinaciones que se auto-compensan y arrojan el mismo valor de función objetivo (Bogaerts & Wouwer, 2004; Pitt & Banga, 2019). En caso de que esta situación se presente, en esta investigación se propone realizar 20 optimizaciones con cotas dilatadas, en un orden de magnitud, en la dirección que indique el valor del parámetro.

5.3.8. Estimación de parámetros

El problema de optimización para estimación de parámetros queda entonces definido por el modelo matemático de descrito por las Ecuaciones 4-1 a 4-7, función objetivo de mínimos cuadrados con normalización por valor medio del "estado" (Ecuación 5-15), integrador `ode15s()` de Matlab® R2018 y optimizador enjambre de partículas sintonizado (Tabla **5-1**). Dada la naturaleza estocástica del optimizador enjambre de partículas, se hace necesario aplicar *bootstraping* o muestreo aleatorio con reemplazo (Johnson, 2010). Esta metodología se traduce en la realización de un número definido de optimizaciones independientes, es decir, que parten de semillas aleatorias diferentes. De esta manera es posible conocer el comportamiento de los parámetros en la superficie no convexa descrita por la función objetivo (Good & Hardin, 2012).

5.3.9. Efecto de identificabilidad estructural

En el capítulo 3 se estableció que el modelo de fermentación diaúxica es localmente identificable con el uso de experimentos que posean condiciones iniciales no nulas de glucosa y xilosa. La afirmación anterior fue puesta a prueba mediante la realización de 45 optimizaciones en el conjunto de datos 1 pretratado. Adicionalmente, los parámetros μ_{xil}^{max}, $K_{S,xil}$, $K_{S,xit}$, $q_x^{max}il$, $K_{i,xil}$ y P_{xit} son globalmente identificables en experimentos que posean únicamente xilosa como sustrato, sí los valores de los parámetros μ_{glu}^{max}, $K_{S,glu}$, q_{glu}^{max}, $K_{i,glu}$ y K_r han sido fijados. Con el objetivo de comprobar esta propiedad de identificabilidad y el efecto de la cantidad de datos en la estimación de parámetros, 45 optimizaciones fueron con los conjuntos de datos 2 a 9 y los conjuntos de datos 8 y 9, todos ellos con pretratamiento (detección de puntos atípicos y suavizado por aproximación polinomial) (MatLab, 2018; Savitzky & Golay, 1964). En los conjuntos de datos 2 a 9 solamente se presenta xilosa como sustrato.

5.3.10. Pruebas estadísticas de estimación y validación

a) Intervalos de confianza y correlación de parámetros

Se dice que los parámetros de un modelo matemático son *estimados*, puesto que, desde un punto de vista estadístico su valor verdadero no puede ser conocido con certeza absoluta (Weise, 1985; Englezos & Kalogerakis, 2000; Soize, 2017). En este sentido, el nivel de incertidumbre asociado al valor numérico de los parámetros puede calcularse mediante la matriz de varianza-covarianza (Cov) como indica la Fórmula 5-16 (Nocedal & Wright, 2006; Myers et al., 2012).

$$Cov = \sigma^2 H^{-1}; \qquad \sigma^2 = \frac{\sum_{i=1}^{n} \mathbf{e}_i^2}{n-p}; \qquad H = FIM = \frac{\partial^2 f}{\partial \boldsymbol{\theta}^2} = \frac{\partial f}{\partial \boldsymbol{\theta}}^{'} \frac{\partial f}{\partial \boldsymbol{\theta}} \tag{5-16}$$

en donde σ^2 corresponde a la varianza del modelo, n número de datos experimentales, p número de parámetros y H Hessiano o matriz de segundas derivadas parciales de los estados

respecto a los parámetros. En términos estadísticos, H se conoce como matriz de información de Fisher (FIM) y explica la cantidad de información que una variable posee sobre los parámetros desconocidos (Van den Bos, 2007).

Las varianzas de los parámetros ($S^2(\boldsymbol{\theta})$) estarán ubicadas en la diagonal principal de la matriz Cov.

$$S^2(\boldsymbol{\theta}) = diag(Cov) \tag{5-17}$$

Los intervalos de confianza para los parámetros puede ser calculados mediante la Fórmula 5-18 (Myers et al., 2012).

$$\boldsymbol{\theta} - Z_{\alpha/2}\sqrt{S^2(\boldsymbol{\theta})} \leq \boldsymbol{\theta} \leq \boldsymbol{\theta} + Z_{\alpha/2}\sqrt{S^2(\boldsymbol{\theta})} \tag{5-18}$$

en donde Z corresponde al valor de la distribución Z con nivel de confianza estadística igual a 100(1-α) %, α corresponde al nivel de significancia. Adicionalmente, el nivel de interacción entre los parámetros puede ser calculado a partir de la matriz Cov mediante la Fórmula 5-19 (Myers et al., 2012).

$$\rho_{i,j} = \frac{Cov_{i,j}}{\sqrt{Cov_{i,i}Cov_{j,j}}} \tag{5-19}$$

en donde ρ es el coeficiente de correlación de Pearson entre los parámetros i y j. Sí $\rho \approx 0$ los parámetros son independientes entre sí. Por el contrario, sí $\rho \approx 1$ los parámetros están complemente correlacionados. La correlación implica que cambios en el valor de un parámetro serán compensados con cambios en uno o más parámetros.

Los intervalos de confianza y matriz de correlación para los parámetros estimados del modelo de fermentación diaúxica para producción de xilitol fueron calculados a través de las Fórmulas 5-16 a 5-19.

b) Indicadores de calidad de predicción

Dependiendo del tipo de datos utilizado se pueden distinguir dos tipo de calidad de predicción. La calidad descriptiva hace referencia a la discrepancia del modelo respecto a información que ya conoce (datos experimentales utilizados en estimación). Por otra parte, la calidad predictiva hace referencia la discrepancia del modelo con información que desconoce (datos experimentales diferentes a los utilizados en estimación). Una aproximación comúnmente utilizada para establecer el grado de discrepancia entre la respuesta del modelo y los datos experimentales corresponde a los índices de ajuste (Barrigón et al., 2012). Algunos de estos son mostrados en la Tabla **5-3**, en donde y corresponde a un dato experimental, x respuesta del modelo, n número de datos experimentales y p número total de parámetros (Holst et al., 1993; Gunay, 2007; Gelman et al., 2019).

Tabla 5-3: Indicadores de discrepancia del modelo.

Indicador	Fórmula	
Suma de cuadrados del error (SSE)	$\sum_{i=1}^{n} \mathbf{e}_i^2$	(5-20)
Error cuadrático Medio (RMSE)	$\sqrt{\frac{\sum_{i=1}^{n} \mathbf{e}_i^2}{n}}$	(5-21)
Error relativo medio (MRE)	$\frac{1}{n}\sum_{i=1}^{n} \frac{\lvert\mathbf{e}_i\rvert}{y_i}$	(5-22)
Bondad de ajuste (GoF)	$\sum_{i=1}^{n} \frac{\mathbf{e}_i^2}{y_i}$	(5-23)
Coeficiente de determinación lineal (R^2)	$1 - \frac{\sum_{i=1}^{n} \mathbf{e}_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	(5-24)
Coeficiente de determinación no lineal (R^2)	$\frac{\sum_{i=1}^{n}(x_i - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \bar{y})^2 + \mathbf{e}_i^2}$	(5-25)
Criterio de información de Akaike (AIC)	$-2ln(SSE) + 2p$	(5-26)
Criterio de información bayesiana (BIC)	$-2ln(SSE) + pln(n)$	(5-27)
Error de predicción final (FPE)	$\left[\frac{SSE}{n-p}\right]^2 \frac{n+p}{n-p}$	(5-28)

c) Intervalos de confianza y correlación del predictor

De forma análoga a los intervalos de confianza de los parámetros estimados, es posible calcular la incertidumbre en la respuesta del modelo matemático para un $\boldsymbol{\theta}$ estimado mediante *intervalos de confianza del predictor*, según la Fórmula 5-29 (Myers et al., 2012).

$$\mathbf{x}(\boldsymbol{\theta}) \pm N_{\alpha/2}\sqrt{\sigma^2\,(1 + d_0\,Cov\,d_0)}; \qquad d_0 = \left.\frac{\partial f}{\partial \boldsymbol{\theta}}\right|_{x=x_0} \tag{5-29}$$

en donde $N_{\alpha/2}$ es el valor de la distribución normal con nivel de confianza $\alpha/2$ y d_0 derivada parcial de la respuesta del modelo respecto a los parámetros evaluada en el punto x_0.

Una prueba adicional que debe realizarse al modelo con sus parámetros estimados corresponde a la prueba de autocorrelación de residuales (Ljung, 1990). Este método analiza si los

residuales o discrepancias entre la salida del modelo y las observaciones experimentales se encuentran correlacionadas entre sí. La prueba de autocorrelación se calcula a través de la Fórmula 5-30.

$$r_{ee}(l) = \frac{1}{n-1}\sum_{i=1}^{n-1}\mathbf{e}(i)\mathbf{e}(i+l); \qquad \frac{n}{r_{ee}(0)^2}\sum_{l=1}^{M}\hat{r}_{ee}(l)^2 \leq X_{\alpha}^2(M) \tag{5-30}$$

en donde r_{ee} corresponde al valor de autocorrelación para los residuales e_i en el retraso l, n cantidad total de residuales y $X_{\alpha}^2(M)$ valor de la distribución chi-cuadrado para M grados de libertad y un nivel de significancia estadística α. Sí la condición de $X_{\alpha}^2(M)$ es satisfecha, entonces los residuales son independientes entre sí para un rango M de datos experimentales. La autocorrelación de la respuesta del modelo matemático de fermentación diaúxica para producción de xilitol con intervalos de confianza al 95 % fue calculada mediante el comando `autocorr()` de Matlab® R2018.

d) Índices de sensibilidad

Los análisis de sensibilidad tienen por objetivo establecer el grado de sensibilidad de la respuesta del modelo a incertidumbre en el valor de un parámetro estimado. Estos análisis pueden dividirse en dos categorías: locales si cuantifican el efecto individual de los parámetros y globales si cuantifican el efecto total sobre la salida del modelo debido a la variación simultánea de todos los parámetros (Sun & Sun, 2015). Dada la ventaja que presentan los análisis de tipo global estos son de uso común, y entre ellos destaca el método de Sobol basado en descomposición de varianza (Sobol, 1993). Una explicación más detallada del método de Sobol y el cálculo de sus índices es presentada en el apéndice C.8. Se requiere entonces del cálculo de dos tipos de índices, los índices individuales e índices totales.

Los índices de sensibilidad individuales cuantifican el efecto de la incertidumbre de un parámetro sobre la respuesta del modelo, por otra parte, los índices de sensibilidad total cuantifican el efecto de la incertidumbre de un parámetro y sus interacciones sobre la respuesta del modelo matemático. No obstante, el cálculo de los índices de sensibilidad individuales de Sobol (o índices "pequeños" de Sobol) es computacionalmente demandante, razón por la cual otros estimadores han sido propuestos (Sobol & Myshetskaya, 2008; Saltelli et al., 2010; Kucherenko et al., 2011, 2012; Kucherenko & Song, 2017). Entre estos estimadores destaca el de Sobol-Myshetzskay también llamado "oracle", el cual presenta una velocidad de convergencia mayor respecto a otros estimadores (Kucherenko & Song, 2017). El estimador "oracle" se presenta en la Fórmula 5-31 (Kucherenko et al., 2011).

$$V_i = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{x}(A) - \overline{\mathbf{x}(A)})(\mathbf{x}(C_i) - \mathbf{x}(B)); \quad V = \frac{\sum \mathbf{x}(A)^2}{N}; \quad S_i = \frac{V_i}{V} \tag{5-31}$$

en donde V_i hace referencia a la varianza del parámetro i, N número total de combinaciones

de parámetros, A y B matrices de muestreo independientes, C_i matriz de muestreo combinada, $\mathbf{x}()$ respuesta del modelo dada la respectiva matriz de muestreo, V varianza total de la respuesta del modelo y S_i índice de sensibilidad individual para el parámetro i. Adicionalmente, el índice de sensibilidad total de Sobol se presenta en la Fórmula 5-32.

$$V_{-i} = \frac{1}{2N}\sum_{k=1}^{N}(\mathbf{x}(A) - \mathbf{x}(C_i))^2; \qquad S_{T_i} = 1 - \frac{V_{-i}}{V} \tag{5-32}$$

en donde V_{-i} es la variación de la respuesta del modelo debido a todos los parámetros excepto el parámetro i y S_{T_i} índice de sensibilidad total de Sobol para el parámetro i. Los índices de sensibilidad son valores puntuales, por tanto, deben calcularse por estado del modelo y punto de tiempo.

Una alternativa al método de Sobol corresponde al método de coeficientes de regresión estandarizados (SRC por sus siglas en inglés) (Morales-Rodriguez et al., 2012; Qian & Mahdi, 2020). En esta aproximación la incertidumbre es cuantificada mediante el ajuste de la salida del modelo a una regresión lineal múltiple, tomando los parámetros como variables explicativas. Los coeficientes de regresión estandarizados son calculados mediante la Fórmula 5-33.

$$\mathbf{x}(A) = a + \sum_{i}^{p} b_i \cdot \theta_i; \qquad \sigma_{\mathbf{x}(A)} = \frac{\sum \mathbf{x}(A)^2}{N}; \qquad \sigma_{\theta_i} = \frac{\sum A_i^2}{N}; \qquad B_i = \frac{\sigma_{\theta_i}}{\sigma_{\mathbf{x}(A)}} \cdot b_i \tag{5-33}$$

en donde $\mathbf{x}(A)$ corresponde a la respuesta del modelo para matriz de muestreo A, a intercepto de la regresión, b_i coeficiente de regresión del parámetro i, $\sigma_{\mathbf{x}(A)}$ varianza de la salida del modelo, $\sigma_{\boldsymbol{\theta}_i}$ varianza del parámetro i y B_i coeficiente de regresión estandarizado del parámetro i. Si $B_i \approx |1|$ el parámetro i presenta una alta contribución a la varianza de la respuesta del modelo, por tanto, un alto nivel de sensibilidad. Si $B_i \approx 0$ el parámetro i no presenta una influencia relevante en la salida del modelo, por tanto, baja sensibilidad.

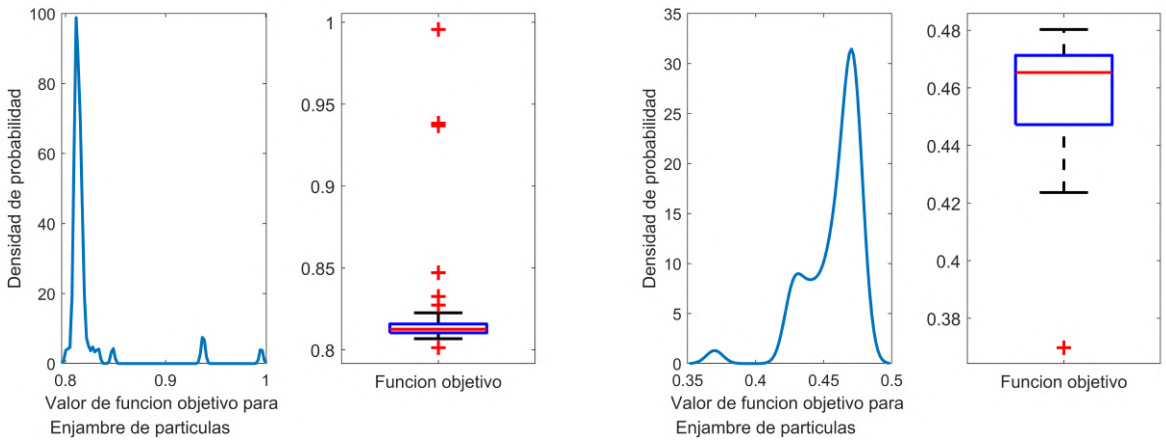
Los perfiles de índices de sensibilidad de Sobol para los parámetros estimados del modelo de fermentación diaúxica para producción de xilitol fueron calculados mediante las Fórmulas 5-31 y 5-32. Las matrices de muestreo A y B fueron generadas para una perturbación del 25 % en el valor de los parámetros y 4000 muestras mediante el comando `lhsdesign` de Matlab® R2018, el cual se basa en una distribución de probabilidad uniforme aplicada al método Latin Hypercube Sampling (LHS). Adicionalmente, nuevos perfiles de índices de sensibilidad fueron calculados con matrices de muestreo alternativas generadas con el comando `lhsnorm` de Matlab® R2018, el cual al basarse en distribuciones de probabilidad normal permite la inclusión de la matriz de varianza-covarianza de los parámetros estimados. A modo de comparación, perfiles de coeficientes de regresión estandarizados fueron calculados con las mismas matrices de muestreo generadas previamente.

5.4. Resultados

A continuación, se presentan los resultados obtenidos para la estimación de parámetros del modelo matemático de fermentación diaúxica para producción de xilitol. En los resultados se incluyen el análisis de los efectos del pretratamiento y cantidad de datos experimentales, características de identificabilidad estructural de los parámetros en la solución del problema de estimación de parámetros, análisis de incertidumbre para los parámetros estimados y su efecto en la capacidad predictiva del modelo.

5.4.1. Tratamiento de datos

La Figura **5-2** presenta la dispersión de valores $\varphi(\boldsymbol{\theta})$ obtenidos para los conjuntos de datos (a) no tratado y (b) tratados. El tratamiento de datos (capítulo 2) fue realizado mediante detección y reemplazo de puntos atípicos junto con eliminación de ruido aleatorio. Las dispersiones de $\varphi(\boldsymbol{\theta})$ para ambos grupos presentan una baja cantidad de *outliers* y un rango intercuartil reducido. Sin embargo, el desempeño obtenido con los datos pretratados es superior alcanzando consistentemente un valor de función objetivo alrededor de 0.46, en comparación con los datos sin pretratar que se encuentran alrededor de 0.81.



(a) Datos sin pretratamiento (b) Datos con pretratamiento

Figura 5-2: Dispersión de valores de función objetivo por tipo de datos.

Las trayectorias del modelo obtenidas con los parámetros estimados con datos experimentales pretratados y no pretratados se presentan en la Figura **C-1**. Se observa similitud de las trayectorias tanto entre tipos de datos como entre estados con valores R^2 promedio de 88.74 % y 87.58 % para datos sin pretratamiento y pretratados, respectivamente. Lo anterior indica que, efectivamente, el pretratamiento de datos experimentales presentó una reducción en el valor de $\varphi(\boldsymbol{\theta})$ debido a la eliminación de puntos atípicos y ruido aleatorio.

La dispersión de valores estimados para los parámetros μ_{glu}^{max} y μ_{xit}^{max} con y sin pretratamiento de datos, se muestran en la Figura **5-3**. El parámetro μ_{glu}^{max} presenta una tendencia respecto a $\varphi(\boldsymbol{\theta})$, disminuyendo su valor conforme se aproxima al mínimo global. Este comportamiento es análogo para los parámetros $K_{S,glu}$, K_r, q_{glu}^{max} (ver Figura **C-2**). En contraste, el parámetro μ_{xit}^{max} presentan aglomeración sobre las cotas, situación que se repite para los parámetros $K_{i,glu}$, $K_{S,xil}$, $K_{S,xit}$, q_{xil}^{max} y $K_{i,xil}$ (ver Figura **C-3**). Sin embargo, el parámetro P_{xit} tiende a un valor constante (ver Figura **C-4**).

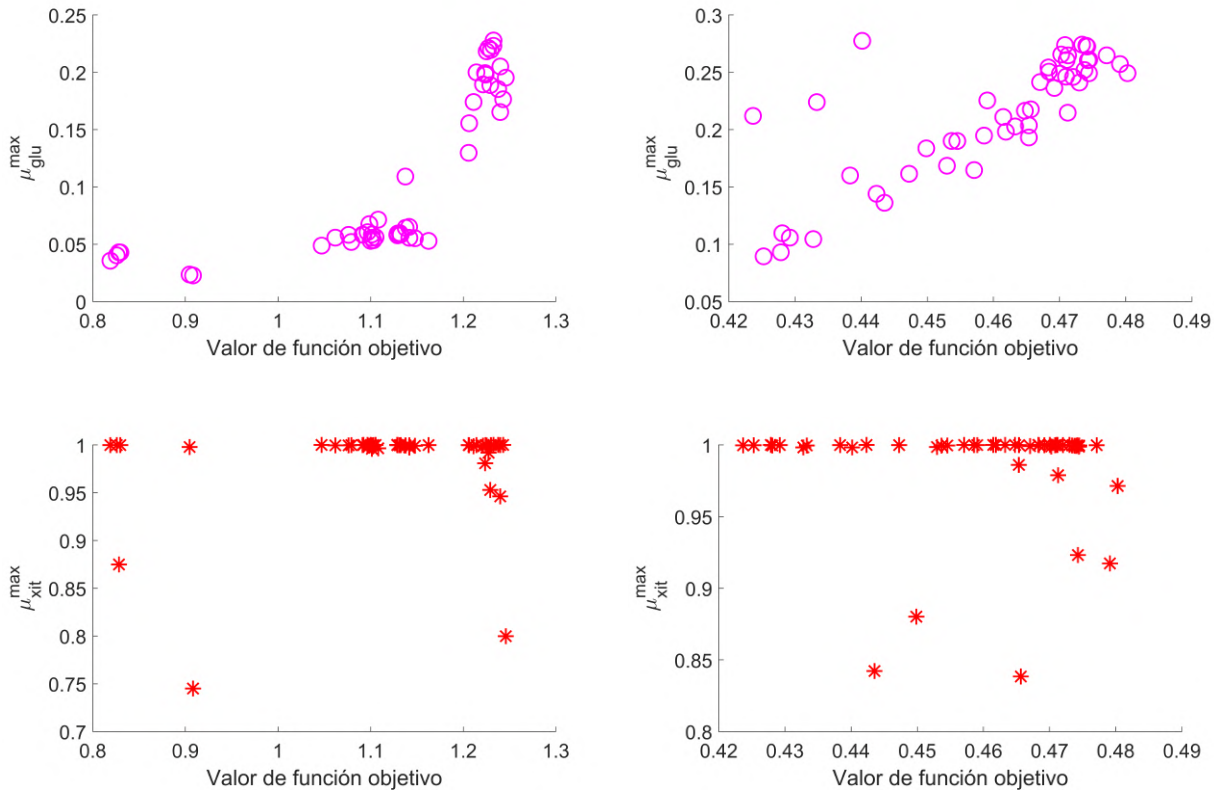


Figura 5-3: Valores de parámetros vs valor de función objetivo para los parámetros μ_{glu}^{max} y μ_{xit}^{max}. Izquierda: datos originales, Derecha: datos pretratados.

Es posible observar una diferencia en los valores estimados entre datos pretratados y originales para aquellos parámetros que presentan una tendencia. Por tanto, el pretratamiento de datos tendría influencia en el valor de $\varphi(\boldsymbol{\theta})$ alcanzado y en el valor estimado de los parámetros. Desde otra perspectiva, la dispersión de valores estimados respecto al valor de $\varphi(\boldsymbol{\theta})$ permite obtener una proyección del comportamiento del valor de un parámetro en la superficie no convexa de la función objetivo. Así, una tendencia definida mostraría que el parámetro es convergente a un valor, y por tanto, sería observable y en principio prácticamente identificable (Kreutz et al., 2013; Kroll et al., 2017). Un procedimiento similar de pretratamiento

de datos ha sido reportado en la estimación de parámetros de un modelo de biorreactor para crecimiento de células de mamífero con 52 especies químicas, 53 reacciones y 31 parámetros (O'Brien et al., 2021).

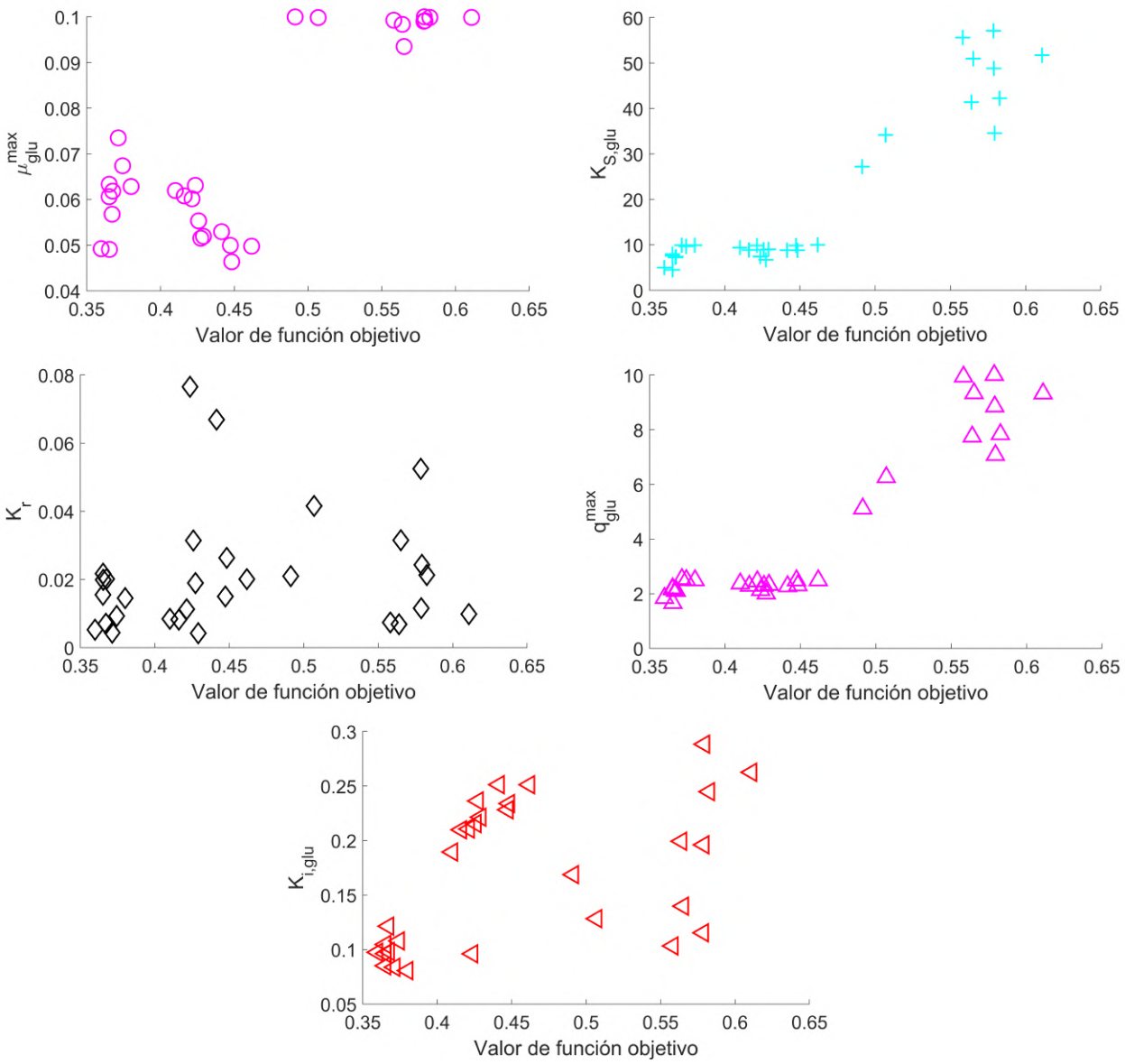


Figura 5-4: Estimación de parámetros μ^{max}_{glu}, $K_{S,glu}$, K_r. q^{max}_{glu}, $K_{i,glu}$ y μ^{max}_{xit} con limites extendidos.

Para descartar el efecto de un límite no satisfactorio en las cotas para los parámetros $K_{i,glu}$, μ^{max}_{xit}, $K_{S,xil}$, $K_{S,xit}$, q^{max}_{xil} y $K_{i,xil}$, las Figuras **5-4** y **5-5** presentan 20 estimaciones de parámetros adicionales con datos experimentales pretratados y cotas extendidas en la dirección requerida. Se observa que los parámetros ya mencionados conservan una dispersión uniforme que varía en por lo menos un orden de magnitud, excepto para $K_{i,glu}$ el cual presenta ten-

dencia a un valor constante. Según lo anterior, la aglomeración de valores sobre la cota para el parámetro $K_{i,glu}$ se debió a un límite no satisfactorio en la Tabla **5-2**. Por otra parte, los parámetros μ_{xit}^{max}, $K_{S,xit}$, q_{xil}^{max} y $K_{i,xil}$ presentan tanto una dispersión uniforme de valores estimados como múltiples valores para un mismo $\varphi(\boldsymbol{\theta})$, lo cual indica alta correlación de parámetros (Villaverde, 2019).

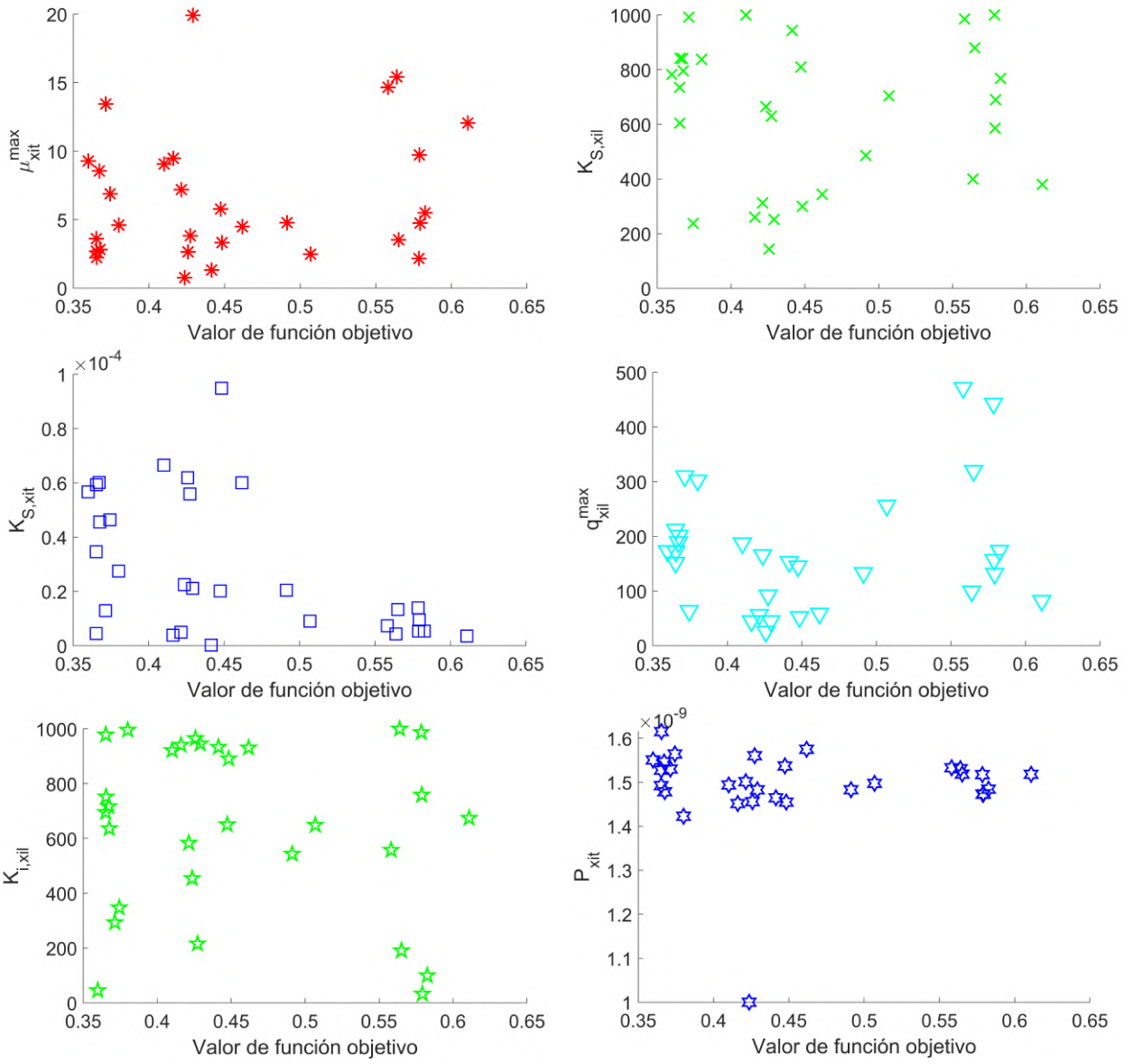


Figura 5-5: Estimación de parámetros μ_{xit}^{max}, $K_{S,xil}$, $K_{S,xit}$, q_{xil}^{max}, $K_{i,xil}$ y P_{xit} con limites extendidos.

El efecto de acumulación sobre las cotas y la interacción entre parámetros observada en las Figuras **5-4** y **5-5** exponen la dificultad que representa la delimitación de Θ. Sin embargo, la aproximación utilizada en este trabajo para establecer Θ es aceptable, al presentar solo

un límite insatisfactorio para el parámetro $K_{i,\,glu}$. Otras metodologías se han propuesto para abordar esta problemática, siendo una de ellas la reducción de Θ basada en funciones de clasificación generadas con máquinas de soporte vectorial, la cual utiliza un muestreo de tipo Monte-Carlo por Cadena de Markov (MCMC por sus siglas en inglés) del espacio Θ (Hasenauer et al., 2010). Otra metodología se basa en muestreo aleatorio del espacio paramétrico Θ y el valor de la función objetivo para reducir la amplitud de las cotas en combinación con optimización regularizada, la cual ha sido aplicada a modelos de osciladores biológicos (Pitt & Banga, 2019). Adicionalmente, esta última aproximación expone dos ventajas, por una parte, no requiere de conocimiento *a priori* de los parámetros y, por otra parte, el término de regularización reduce los efectos de interacción entre parámetros (Hassan et al., 2013; Wang & Wang, 2014).

5.4.2. Identificabilidad práctica vs identificabilidad estructural

Desde la perspectiva de identificabilidad estructural, los 11 parámetros del modelo son localmente identificables bajo las condiciones experimentales del conjunto de datos 1. Sin embargo, los resultados anteriores indican que únicamente los parámetros μ_{glu}^{max}, $K_{S,\,glu}$, q_{glu}^{max}, K_r y $K_{i,\,glu}$ relacionados con glucosa y P_{xit} relacionado con xilitol, serían prácticamente identificables. Por otra parte, los parámetros μ_{xit}^{max}, $K_{S,\,xil}$, $K_{S,\,xit}$, q_{xil}^{max} y $K_{i,\,xil}$, relacionados a xilosa y xilitol serían *prácticamente no identificables*. Tomando en consideración estos resultados, es posible afirmar que sí bien la identificabilidad estructural es una **condición necesaria** al momento de estimar los parámetros de un modelo matemático, **no es suficiente**, puesto que no puede saberse *a priori* la cantidad de información experimental requerida para lograr la identificabilidad práctica completa de $\boldsymbol{\theta}$. Esta afirmación tiene mayor relevancia al tomar en cuenta que el pretratamiento de datos experimentales tiene por objetivo acercar la información experimental a su forma "perfecta" (sin error experimental), suposición utilizada en los análisis de identificabilidad estructural. Esto indica que no solo basta con que el diseño del experimento lleve a la identificabilidad estructural del modelo, también es requerida una cantidad mínima de experimentos con dicha característica (Villaverde et al., 2019). Adicionalmente, puede utilizarse diseño óptimo de experimentos (DOE) para mejorar la información contenida en los datos experimentales (Moser et al., 2021).

Bajo una condición inicial que solo posea xilosa como sustrato, el análisis de identificabilidad estructural determinó que los parámetros μ_{xit}^{max}, q_{xil}^{max}, $K_{S,\,xil}$, $K_{S,\,xit}$ y P_{xit} son estructuralmente globalmente identificables, sí el valor de μ_{glu}^{max}, $K_{S,\,glu}$, q_{glu}^{max}, K_r, $K_{i,\,glu}$ y $K_{i,\,xil}$ es fijado. Es decir, en principio el uso de un experimento con esta característica para el modelo de caso de estudio permitiría siempre encontrar un valor único para estos parámetros. La Tabla **5-4** presenta el valor fijado para los parámetros μ_{glu}^{max}, $K_{S,\,glu}$, q_{glu}^{max}, K_r y $K_{i,\,glu}$ correspondientes al valor alcanzado en el mínimo global de $\varphi(\boldsymbol{\theta})$. La Tabla **5-5** presenta los valores de los parámetros μ_{xit}^{max}, q_{xil}^{max}, $K_{S,\,xil}$, $K_{S,\,xit}$, $K_{i,\,xil}$ y P_{xit} estimados con los conjuntos de datos 2

a 9, y los conjuntos de datos 8 y 9. El parámetro $K_{i,\,xil}$ fue re-estimado al no presentar un valor único en el mínimo global de $\varphi(\boldsymbol{\theta})$.

Tabla 5-4: Parámetros fijados, estimados con el conjunto de datos 1.

Parámetro	Valor	Parámetro	Valor
μ_{glu}^{max}	6.183x10^{-2}	$K_{i,\,glu}$	9.747x10^{-2}
$K_{S,\,glu}$	7.5586	K_r	1.452x10^{-2}
q_{glu}^{max}	2.1546	-	-

Tabla 5-5: Parámetros estimados con los conjuntos de datos 2 a 9 y conjuntos 8 y 9.

Parámetro	Conjuntos de datos 2 a 9	Conjuntos de datos 8 y 9
μ_{xit}^{max}	4.458x10^{-3}	3.73x10^{-2}
q_{xil}^{max}	8.284x10^{-2}	7.939x10^{-2}
$K_{S,\,xil}$	5.909x10^{-3}	8.6672
$K_{S,\,xit}$	5.116x10^{-5}	99.999
$K_{i,\,xil}$	19.318	99.999
P_{xit}	7.823x10^{-11}	5.352x10^{-9}

Los parámetros obtenidos con los diferentes conjuntos de datos presentan diferencias en mínimo un orden de magnitud, a excepción de q_{xil}^{max} el cual es similar en ambos casos. La diferencia en el valor de los parámetros puede ser atribuida a una alta interacción o correlación entre parámetros como consecuencia de insuficiencia en cantidad y calidad de información experimental en relación al número de parámetros a identificar. Estos aspectos se encuentran enlazados dado que, en general, cuando existen parámetros correlacionados en un modelo matemático se producirán desviaciones de los valores "verdaderos" (en el sentido estadístico) de los parámetros, que serán directamente proporcionales al nivel de correlación e inversamente proporcionales a la cantidad de datos (Shieh & Fouladi, 2003).

Dicho fenómeno es apreciable en los parámetros estimados con los conjuntos de datos 8 y 9 (160 observaciones experimentales), los cuales presentan magnitudes mayores y 2 parámetros prácticamente no identificables ($K_{S,\,xit}$ y $K_{S,\,xil}$) en comparación con los valores estimados obtenidos con los conjuntos 2 a 9 (348 observaciones experimentales). De forma particular, el valor elevado de $K_{S,\,xit}$ y $K_{S,\,xil}$ es una señal de interacción al tomar en cuenta que hacen parte de un término racional (e.g. un aumento en el numerador puede ser compensado por un aumento en el denominador). Adicionalmente y como ejercicio académico, se determinó la relevancia de aumentar la cantidad de observaciones para un mismo experimento a través de la re-estimación de los 11 parámetros del modelo matemático con el conjunto de datos 1 aumentado de 44 a 484 datos experimentales mediante el comando `spline()` de

Matlab® R2018. Estos resultados se presentan en el apéndice C.4. Se confirmó que un aumento artificial en la cantidad de observaciones experimentales no cambia las propiedades de identificabilidad práctica para un mismo conjunto de datos, por lo tanto, se requiere mayor información y no solo cantidad. En el caso de modelos matemáticos basados en ecuaciones diferenciales ordinarias, nueva información puede ser obtenida a través de nuevas trayectorias, lo que implica experimentos con diferentes condiciones iniciales.

5.4.3. Calidad de la estimación de parámetros

Los parámetros estimados del modelo de fermentación diaúxica para producción de xilitol junto con sus intervalos de confianza obtenidos con los conjuntos de datos experimentales 1 a 9 son mostrados en la Tabla **5-6**. Los parámetros estimados presentaron un coeficiente de variación inferior al 100 %, a excepción de $K_{i,\,glu}$, por tanto, se alcanzó identificabilidad práctica para 10 de los 11 parámetros. No obstante, es posible agrupar los parámetros prácticamente identificables con alta, media y baja incertidumbre. El grupo de parámetros con alta incertidumbre, superior al 80 %, contendría a μ_{glu}^{max} y K_r. Este grado de incertidumbre se atribuye a la limitada cantidad de información experimental relacionada con el consumo de glucosa (22 observaciones experimentales de glucosa y biomasa). Por otra parte, el grupo de parámetros con incertidumbre media, cercana al 50 %, estaría formado por μ_{xit}^{max}, $K_{S,\,xil}$ y $K_{S,\,xit}$, cuyo nivel de incertidumbre podría atribuirse a una elevada correlación, especialmente para μ_{xit}^{max} y $K_{S,\,xit}$ los cuales hacen parte del mismo término racional. El grupo de parámetros con baja incertidumbre esta conformado por $K_{S,\,glu}$, q_{glu}^{max}, q_{xil}^{max}, $K_{i,\,xil}$ y P_{xit}. Dado que 4 de los 5 parámetros de este grupo están asociados a consumo de xilosa y producción de xilitol, su baja incertidumbre puede ser atribuida a la cantidad y calidad de la información experimental (370 observaciones experimentales con diferentes condiciones iniciales), así mismo, se esperaría baja correlación para estos parámetros. La falta de identificabilidad práctica de $K_{i,\,glu}$, que representa la inhibición del transporte de xilosa debido a la presencia de glucosa en el medio de cultivo, puede ser atribuida una reducida cantidad de información experimental, puesto que solo el conjunto de datos 1 presentaba simultáneamente concentraciones iniciales de glucosa y xilosa.

En cuanto a los valores estimados de los parámetros, en esta investigación se encuentran diferencias con respecto a otros autores. Los parámetros calculados por Tochampa et al. son en general mayores por uno o dos órdenes de magnitud, lo cual podría ser debido a el uso de diferentes conjuntos de datos experimentales Tochampa et al. (2015). Adicionalmente, la situación anterior se repite en la investigación realizada por (Prado-Rubio et al., 2015), en la cual se utilizaron los mismos datos experimentales. Esto indica que las diferencias en magnitud entre los valores estimados de los parámetros en la Tabla **5-6** podrían ser atribuidas a alta interacción entre parámetros.

Tabla 5-6: Estimados, desviación estándar, intervalos de confianza al 95 % y coeficiente de variación para los parámetros del modelo de fermentación diaúxica de glucosa y xilosa para producción de xilitol.

Parámetros	Estimado	Desviación estándar	Intervalo de confianza	Coeficiente de variación
μ_{glu}^{max}	$6.183\text{x}10^{-2}$	$3.23\text{x}10^{-2}$	$\pm\ 5.32\text{x}10^{-2}$	52.28 %
$K_{S,\,glu}$	7.5586	$2.15\text{x}10^{-3}$	$\pm\ 3.54\text{x}10^{-3}$	0.03 %
q_{glu}^{max}	2.1546	$1.69\text{x}10^{-2}$	$\pm\ 2.78\text{x}10^{-2}$	0.78 %
$K_{i,\,glu}$	$9.747\text{x}10^{-2}$	0.48	$\pm$ 0.798	497 %
K_r	$1.452\text{x}10^{-2}$	$8.50\text{x}10^{-3}$	$\pm\ 1.39\text{x}10^{-2}$	58.57 %
μ_{xit}^{max}	$4.458\text{x}10^{-3}$	$1.37\text{x}10^{-3}$	$\pm\ 2.26\text{x}10^{-3}$	30.89 %
$K_{S,\,xil}$	$5.909\text{x}10^{-3}$	$2.64\text{x}10^{-3}$	$\pm\ 4.34\text{x}10^{-3}$	44.69 %
$K_{S,\,xit}$	$5.116\text{x}10^{-5}$	$1.21\text{x}10^{-5}$	$\pm\ 1.99\text{x}10^{-5}$	23.69 %
q_{xil}^{max}	$8.284\text{x}10^{-2}$	$2.10\text{x}10^{-4}$	$\pm\ 3.47\text{x}10^{-4}$	0.25 %
$K_{i,\,xil}$	19.318	0.17	$\pm$ 0.28	0.87 %
P_{xit}	$7.823\text{x}10^{-11}$	$1.22\text{x}10^{-13}$	$\pm\ 2.02^{-13}$	0.16 %

La correlación entre estos parámetros se muestra como mapa de calor en la Figura **5-6**. En términos generales, se presentan 6 correlaciones significativas las cuales se especifican en la Tabla **C-2**. Adicionalmente, la Figura **C-7** presenta el número de interacciones en las que participa cada parámetro. μ_{glu}^{max} y P_{xit} no presentaron interacciones, por lo tanto, la elevada incertidumbre presente en μ_{glu}^{max} se debe efectivamente una reducida información experimental. Los parámetros $K_{S,\,glu}$, $K_{S,\,xil}$, q_{glu}^{max}, q_{xil}^{max}, $K_{i,\,glu}$ y $K_{i,\,xit}$ presentaron 1 interacción, lo cual sumado a una elevada cantidad de información experimental explica la baja incertidumbre calculada para $K_{S,\,glu}$, q_{glu}^{max}, q_{xil}^{max} y $K_{i,\,xit}$. En contraste, los parámetros μ_{xit}^{max}, $K_{S,\,xit}$ y K_r presentan la mayor cantidad de interacciones, formando inclusive una correlación triple de tipo $\mu_{xit}^{max}(\uparrow)$ $K_{S,\,xit}(\downarrow)$ $K_r(\downarrow)$. Lo anterior explicaría la elevada incertidumbre calculada para K_r y $K_{S,\,xit}$.

La correlación entre los parámetros del modelo matemático analizado refleja tanto el carácter altamente no lineal del modelo como el impacto negativo de un diseño experimental no adecuado. En este caso particular, sería requerido un mayor número de experimentos que presentaran ambos sustratos con diferentes condiciones iniciales, lo cual nuevamente se conecta con el resultado del análisis de identificabilidad estructural. No obstante, si pudiera generarse la suficiente información experimental podrían reducirse los intervalos de confianza de los parámetros a través del "Diseño Óptimo de Experimentos" (OED por sus siglas en inglés) (Abt et al., 2018).

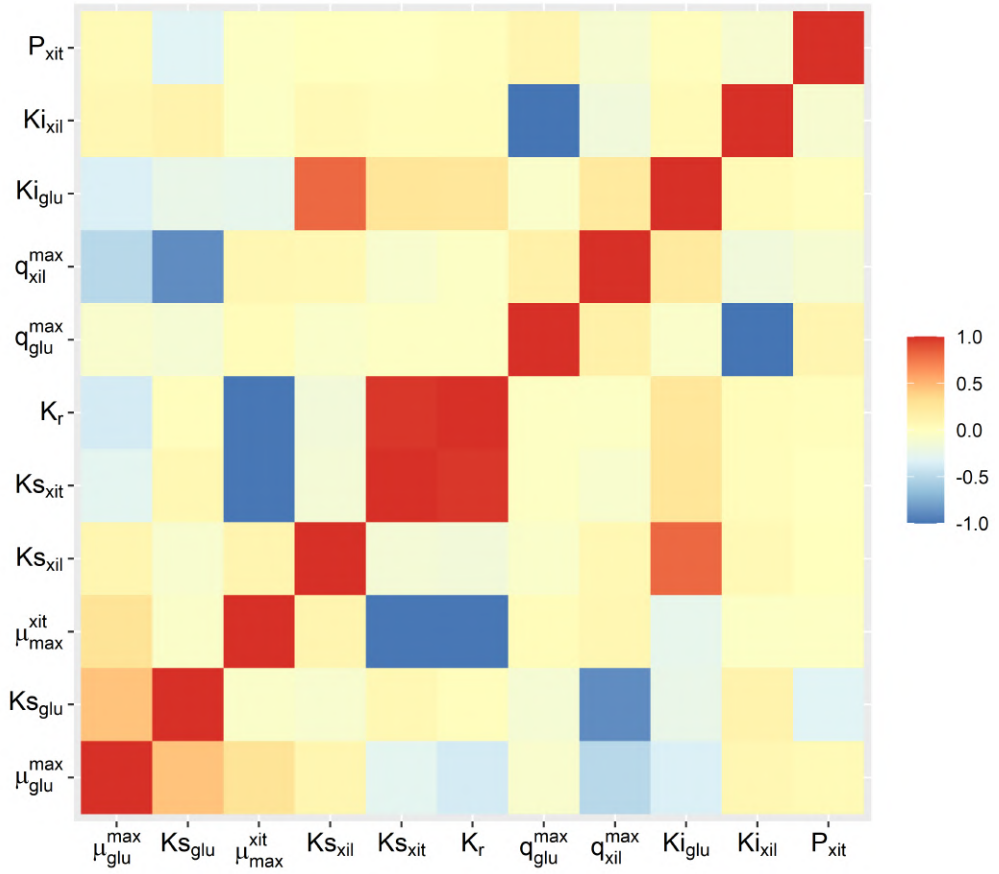


Figura 5-6: Matriz de correlación de los parámetros del modelo de fermentación diaúxica de glucosa y xilosa para producción de xilitol.

Desde perspectiva de OED, la matriz de información de Fisher puede ser utilizada como función objetivo con diferentes efectos como la minimización de la incertidumbre de parámetros (Criterio D: $\max Det(FIM)$), minimización de la varianza de los parámetros (Criterio A: $\max traza(FIM)$), minimización del intervalo de confianza más amplio (Criterio E: $\max \lambda_{min}(FIM)$), minimización de la correlación entre parámetros (Criterio E modificado: $\min \lambda_{max}(FIM)/\lambda_{min}(FIM)$) o minimización de la mayor incertidumbre (Criterio M: $\max min(diag(FIM))$) (García et al., 2017). Estos criterios han sido aplicados en la estimación de parámetros de modelos de fermentación fed-batch y Lokta-Volterra, encontrándose que el criterio E presenta los mejores resultados (Telen et al., 2012). Otros trabajos incluyen la estimación de parámetros de un modelo dinámico de expresión genética para un gen integrado (Braniff et al., 2019).

5.4.4. Indicadores de ajuste

Dado que los indicadores son función de los residuales del modelo, un menor valor del indicador señala mayor cercanía de la respuesta del modelo a los datos experimentales. En general,

si el valor del indicador calculado con los conjuntos experimentales de validación es similar a aquel obtenido en estimación, entonces podría afirmarse que el modelo presenta capacidad predictiva. Sin embargo, el valor de un indicador puede presentar diferencias debido no a la capacidad predictiva del modelo, sino a otros factores como diferencias en la cantidad o magnitud de datos experimentales.

La Tabla **5-7** presenta los indicadores de desempeño calculados tanto para estimación como validación con el modelo de caso de estudio y los parámetros mostrados en la Tabla **5-6**. Se observa que, en general, los indicadores presentan un mejor ajuste en estimación respecto a validación. Específicamente, el indicador SSE presenta la mayor diferencia en magnitud, no obstante, este indicador no presenta ningún tipo de normalización por lo que la diferencia puede ser atribuible a la diferencia en las magnitudes de los datos experimentales y no a su cantidad (392 datos experimentales en estimación vs 444 datos experimentales en validación). Esta situación es aplicable además a los indicadores RMSE y FPE. Por otra parte, los indicadores MRE y GoF que incluyen normalización por valor experimental presentan mayor similitud en sus valores para estimación y validación, lo que implica que la diferencia en los indicadores SSE, RMSE y FPE es debida a una mayor magnitud de los datos experimentales de validación. Con respecto a la selección de los conjuntos de datos para estimación y validación, se tuvo en consideración la naturaleza de los experimentos de tipo batch repetido. Se esperaría inicialmente que el modelo presente una calibración satisfactoria con una fracción de los experimentos y pueda predecir los restantes al ser del mismo tipo.

Un caso particular se da para los indicadores de Criterio de información de Akaike (AIC) y Bayesiana (BIC), los cuales consideran tanto el ajuste como la cantidad de parámetros del modelo matemático. Sin embargo, dado que los datos experimentales afectan directamente a la función objetivo cuando se requiere de normalización (y en consecuencia al valor de los parámetros), el uso de los indicadores AIC y BIC no sería recomendable para comparación entre diferentes conjuntos de datos experimentales. Por otra parte, la sensibilidad de los indicadores a las características de los conjuntos experimentales hace necesaria la consideración de estas diferencias al momento de seleccionar un indicador.

En caso de que los conjuntos de datos presenten similitud en número de datos experimentales y escalas, los indicadores SSE, RMSE y FPE brindarían una comparación adecuada de la capacidad predictiva del modelo. Por otra parte, si se presentan diferencias entre los conjuntos de datos los indicadores MRE y GoF arrojarían una mejor comparación. En caso de que se comparen diferentes modelos con un ajuste similar, los criterios AIK y AIB serían la opción a elegir. Entre las aplicaciones de estos indicadores se incluye el trabajo realizado por Urniezius y Survyla, en donde diversos indicadores de ajuste fueron cuantificados para comparar el desempeño de un modelo de producción de proteínas en la bacteria *Escherichia coli* (Urniezius & Survyla, 2019). Se encontró similitud en el valor de los indicadores para

los mismos conjuntos de datos.

Tabla 5-7: Indicadores de desempeño con parámetros estimados.

Indicador	Estimación	Validación
SSE	12945.89	81445.7
RMSE	5.7468	13.5438
MRE	2.2406	1.4724
GoF	1610	2230
AIC	3.0629	-0.6154
BIC	46.7468	44.4387
FPE	35.9407	197.6521

Un indicador de ajuste ampliamente utilizado es el coeficiente de determinación o R^2, el cual cuantifica el porcentaje de varianza de los datos experimentales explicada por el modelo matemático (Quinn & Keough, 2002). Sin embargo, este indicador (definido para modelos lineales) es comúnmente aplicado en modelos altamente no lineales, aunque alternativas no lineales han sido propuestas (Gelman et al., 2019). La Tabla **C-3** presenta la comparación entre los indices R^2 lineal (R^2_{lineal}) y no lineal ($R^2_{no\,lineal}$) por estado y conjunto de datos. Se encontró que $R^2_{no\,lineal}$ presenta baja sensibilidad en comparación con R^2_{lineal}.

El indicador $R^2_{no\,lineal}$ presenta la ventaja de no arrojar valores negativos, sin embargo, a pesar de que el modelo no se ajuste a los datos experimentales su valor es de mínimo 0.5. Esto se evidencia con el ajuste promedio para biomasa en estimación, en donde se presentan valores de R^2_{lineal} y $R^2_{no\,lineal}$ de -2.813 y 0.548, respectivamente. Debido a lo anterior, es preferible el uso del indicador R^2_{lineal} por su mayor sensibilidad y claridad de interpretación. En promedio se presentó un mejor ajuste del modelo para los estados de xilosa y xilitol, tanto en estimación (0.628 y 0.901) como en validación (0.138 y 0.653) respecto a biomasa, el cual presenta valores negativos para estimación y validación de -2.813 y -13.35, respectivamente.

5.4.5. Incertidumbre en la respuesta del modelo

La Figura **5-7** muestra los intervalos de confianza obtenidos para el conjunto de datos 1 a 9 utilizados en la estimación de parámetros. En general, la predicción realizada por el modelo describe los datos experimentales de manera satisfactoria con intervalos de confianza que contienen a los datos experimentales. Sin embargo, se observan discrepancias en los últimos dos conjuntos de datos, especialmente en las concentraciones de biomasa y xilosa. Por otra parte, las concentraciones de xilosa y xilitol exhiben intervalos de confianza estrechos, lo que indica alta certeza en la calibración del modelo. Dado que los intervalos de confianza son calculados a partir de la matriz FIM, la amplitud del intervalo es un reflejo de la cantidad de información experimental capturada por el modelo matemático de fermentación diaúxica

para producción de xilitol. La discrepancia observada en los conjuntos de datos 8 y 9 podría ser atribuida a la adaptación del microorganismo al consumo consecutivo de concentraciones ascendentes de xilosa en cada etapa batch, producto del diseño experimental utilizado (Sirisansaneeyakul et al., 2013).

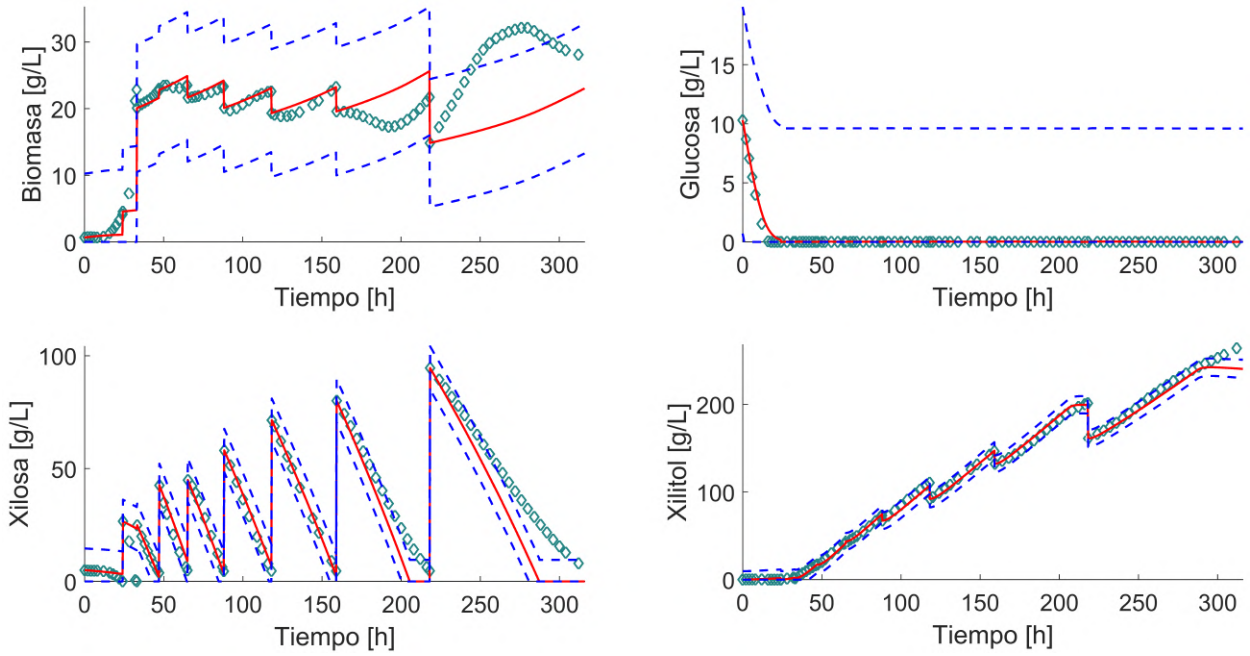


Figura 5-7: Predicción del modelo con intervalos de confianza (95 %) conjuntos de datos 1 a 9. — predicción, - - intervalo de confianza, ◇ dato experimental.

Otra forma de determinar si los parámetros estimados han capturado la información disponible en las observaciones experimentales es a través de la prueba de autocorrelación. La Figura **5-8** presenta los resultados de la prueba de autocorrelación para los residuales del modelo en los conjuntos de datos 1 a 9. Los cuatro estados descritos por el modelo presentan autocorrelaciones superiores a un nivel de confianza del 95 % en retrasos diferentes a cero, lo que indica que el modelo falla al intentar capturar el comportamiento total del sistema. La mayor autocorrelación se presenta en la concentración de biomasa, no obstante, esta variable presenta la mayor incertidumbre experimental, por tanto, no es posible determinar si la discrepancia es debida a la estructura del modelo, error en los datos experimentales o a una combinación de ambas.

La capacidad predictiva del modelo puede conocerse a través de la comparación de la respuesta del modelo con conjuntos de datos diferentes a los utilizados en la estimación de parámetros. Este proceso es conocido como validación de modelos. La Figura **5-9** presenta los intervalos obtenidos para la respuesta del modelo en los conjuntos de datos experimentales 10 a 22. Se observa una mejor predicción de la concentración de biomasa con intervalos de

confianza que contienen en su totalidad a los datos experimentales. Sin embargo, se presenta una subestimación consistente en el consumo de xilosa con intervalos de confianza que no contienen a los datos experimentales y predicción con menor concentración de xilitol al final de cada ciclo batch. Este fenómeno es consistente con lo sucedido en los conjuntos de datos 8 y 9, por lo que podría ser efectivamente atribuido a una adaptación ocurrida en el microorganismo. Tomando en consideración este comportamiento, el microorganismo disminuyó la velocidad de consumo de xilosa y aumento la producción de xilitol, lo que implicaría cambios en los parámetros q_{xil}^{max}, $K_{S,xil}$, P_{xit} y posiblemente μ_{xit}^{max} y $K_{S,xit}$.

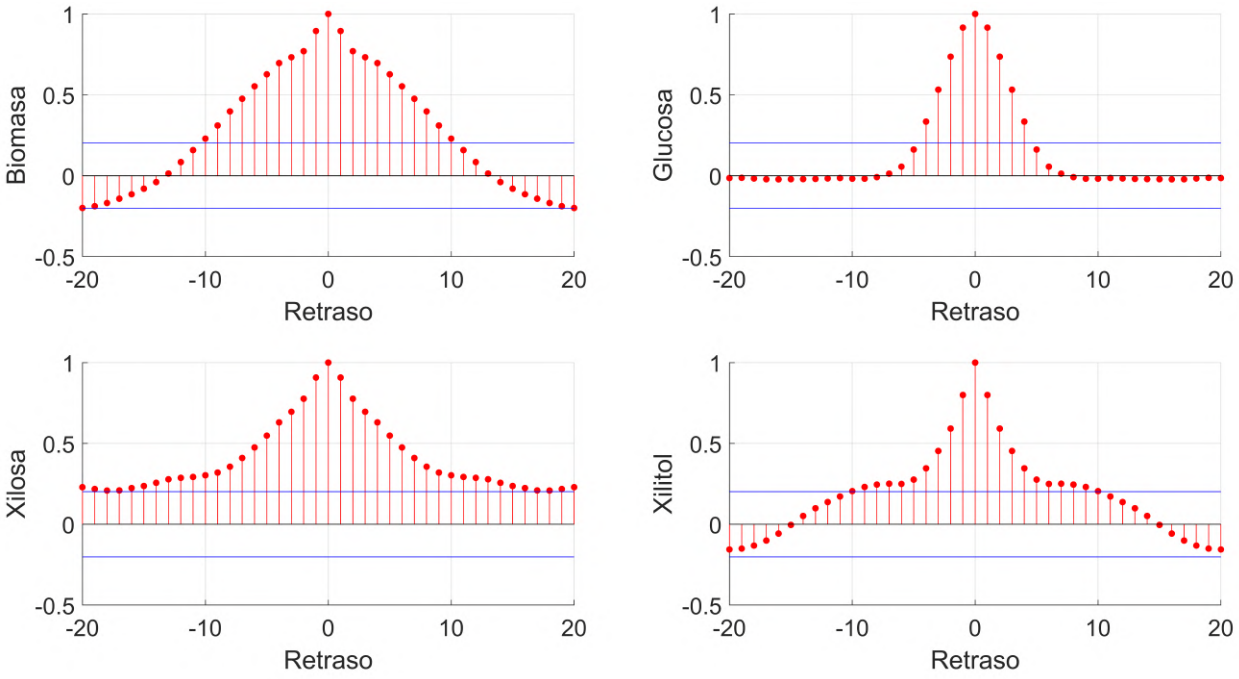


Figura 5-8: Análisis de autocorrelación para los residuales de los conjuntos de datos 1 a 9. (— intervalos de confianza, $\alpha = 95\,\%$).

La adaptación de microorganismos a su medio puede verse reflejada en un modelo matemático a través de cambios en el valor de los parámetros. Esta situación ha sido estudiada previamente para un modelo matemático basado en red metabólica del microorganismo *Saccharomyces cerevisiae* con control de tipo feedback implementado en Matlab, en donde parámetros variables regulan los fluxes metabólicos (Sainz et al., 2003). Así mismo, se ha estudiado la importancia de la precisión del muestreo para control de la tasa de crecimiento específica de un microorganismo y su relevancia en procesos biotecnológicos (Schuler & Marison, 2012). En otro estudio se analizó el efecto de adaptaciones rápidas del microorganismo *S. cerevisiae* en fermentación simultánea y cofermentación de glucosa y xilosa para producción de etanol, encontrando que las adaptaciones ocasionan cambios en la regulación genética que mejoran el rendimiento de etanol (Nielsen et al., 2015).

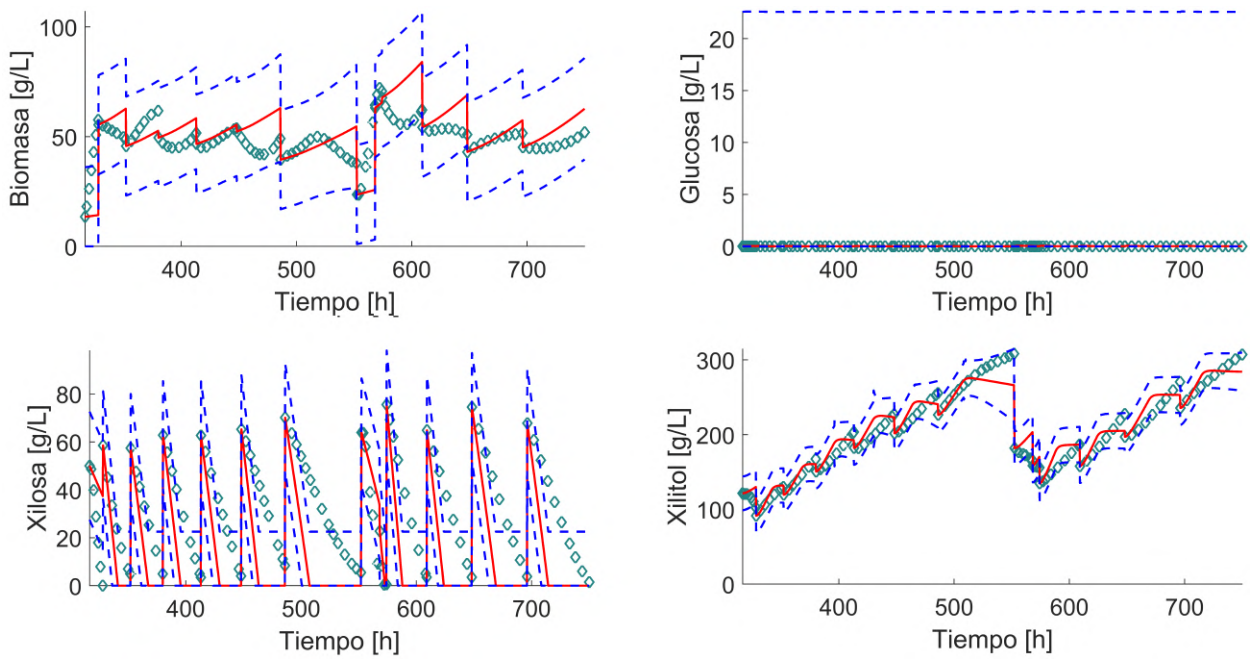


Figura 5-9: Predicción del modelo con intervalos de confianza (95 %) conjuntos de datos 10 a 22. — predicción, - - intervalo de confianza, ◇ dato experimental.

De forma adicional, el análisis de autocorrelación para el conjunto de datos experimentales de validación se presenta en la Figura **C-8**. En contraste con lo ocurrido para los conjuntos de datos utilizados en calibración, el análisis de autocorrelación de los conjuntos de datos de validación indica una mejor aproximación por parte del modelo al comportamiento del sistema. Si bien los cuatro estados exhiben autocorrelación superior al intervalo de confianza del 95 %, la cantidad de retrasos en donde se presenta es notablemente inferior. Esta situación es contradictoria con la Figura **5-9**, en donde se observa que el modelo falla en la predicción de la concentración de xilosa. A manera de hipótesis, el resultado de autocorrelación podría indicar que la estructura del modelo describe correctamente la fenomenología del sistema y la discrepancia es debida netamente al valor de los parámetros estimados. Los análisis de autocorrelación han sido utilizados para compensar los errores de predicción en modelos con problemas estructurales (Villez et al., 2020).

5.4.6. Índices de sensibilidad

La Figura **5-10** presenta de forma detallada los perfiles de índices individuales de Sobol para los 11 parámetros del modelo estudiado, calculados con matriz de muestreo uniforme en los conjuntos de datos 1 a 9. El conjunto de datos 1 presenta un comportamiento diferente respecto a los demás conjuntos de datos experimentales, en donde el modelo es más sensible

a μ_{glu}^{max} y presentan igual sensibilidad $K_{S,glu}$, μ_{xit}^{max}, $K_{S,xil}$, $K_{S,xit}$ y K_r. Sin embargo, esto no es evidente en la figura debido a la superposición de sus perfiles. Por tanto, en presencia de ambos sustratos el modelo es sensible a los parámetros de crecimiento de xilosa y glucosa, pero no a los relacionados con inhibición o transporte.

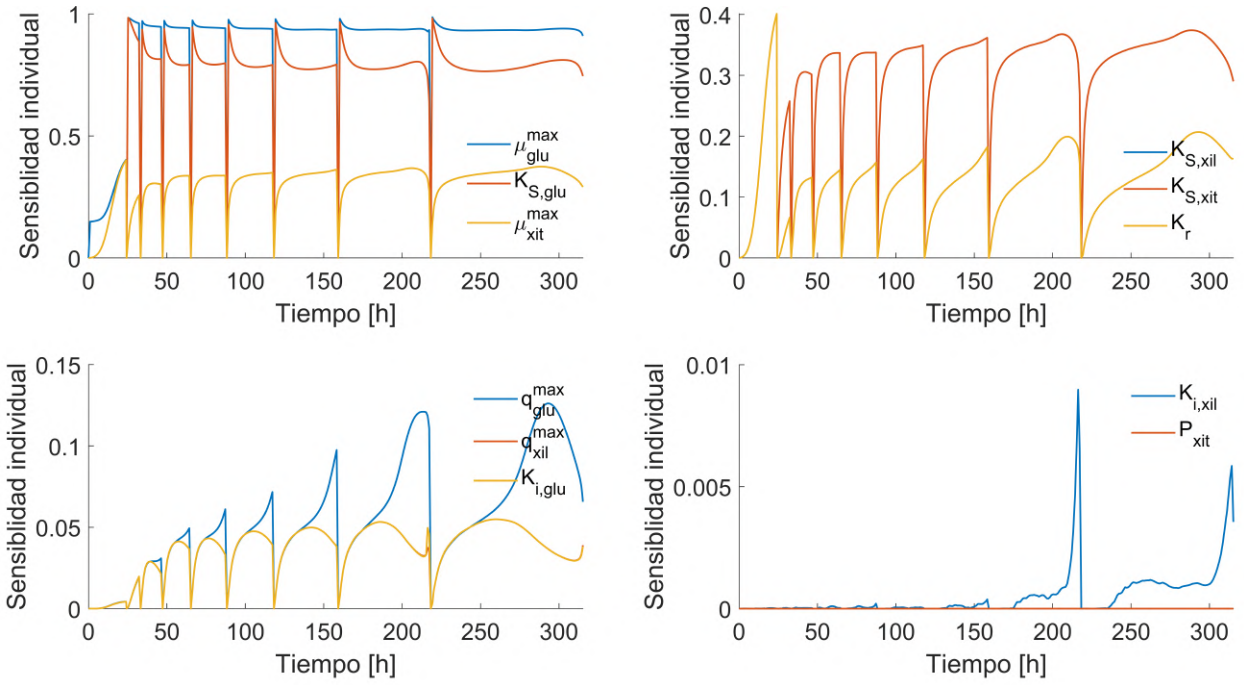


Figura 5-10: Perfiles de índices de sensibilidad individuales de Sobol detallados con distribución uniforme para el modelo matemático de fermentación diaúxica para producción de xilitol.

Dado que los parámetros del modelo poseen significado físico, sus perfiles de sensibilidad podrían brindar información de los fenómenos relevantes para el sistema en un momento determinado. Por otra parte, los conjuntos de datos 2 a 9 presentan un comportamiento similar. En orden descendente de sensibilidad se tienen μ_{glu}^{max}, $K_{S,glu}$, el grupo μ_{xit}^{max}, $K_{S,xil}$ y $K_{S,xit}$ y el grupo q_{glu}^{max}, q_{xil}^{max} y $K_{i,glu}$. Los grupos de parámetros presentan superposición en sus perfiles. A nivel fenomenológico, los perfiles indicarían nuevamente que el transporte y la inhibición no son relevantes. En este caso, el modelo es más sensible a los parámetros relacionados con glucosa a pesar de no estar presente en estos experimentos, lo cual es atribuible al efecto de la correlación de parámetros.

Para el presente caso de estudio se observa que en un mismo instante de tiempo la sumatoria de los índices individuales es superior a 1, esto puede llegar a ocurrir en el caso que los parámetros se encuentren altamente correlacionados (Kucherenko et al., 2012). Debido a la interacción entre parámetros, los índices individuales tendrían entonces tres compo-

nentes: contribución correlacionada por términos de interacción, contribución correlacionada por términos univariados y contribución no correlacionada por términos univariados (Wei et al., 2015). De forma adicional, la Figura **C-9** presenta los perfiles de índices individuales de Sobol por estado, sin embargo, el comportamiento de los perfiles es idéntico entre estados.

La Figura **5-11** presenta los perfiles índices totales de Sobol por parámetro. Dado que los índices totales consideran solo el efecto del parámetro y sus interacciones en la respuesta del modelo, generalmente su magnitud es menor en caso de que los parámetros se encuentren altamente correlacionados (Wei et al., 2015). En este caso, el orden de sensibilidad ya descrito en los perfiles de los índices individuales se mantiene tanto para el conjunto de datos 1 como los conjuntos de datos 2 a 9. Adicionalmente, la Figura **C-10** presenta los perfiles de índices totales de Sobol por estado, sin embargo, el comportamiento de los perfiles es similares para todos los estados.

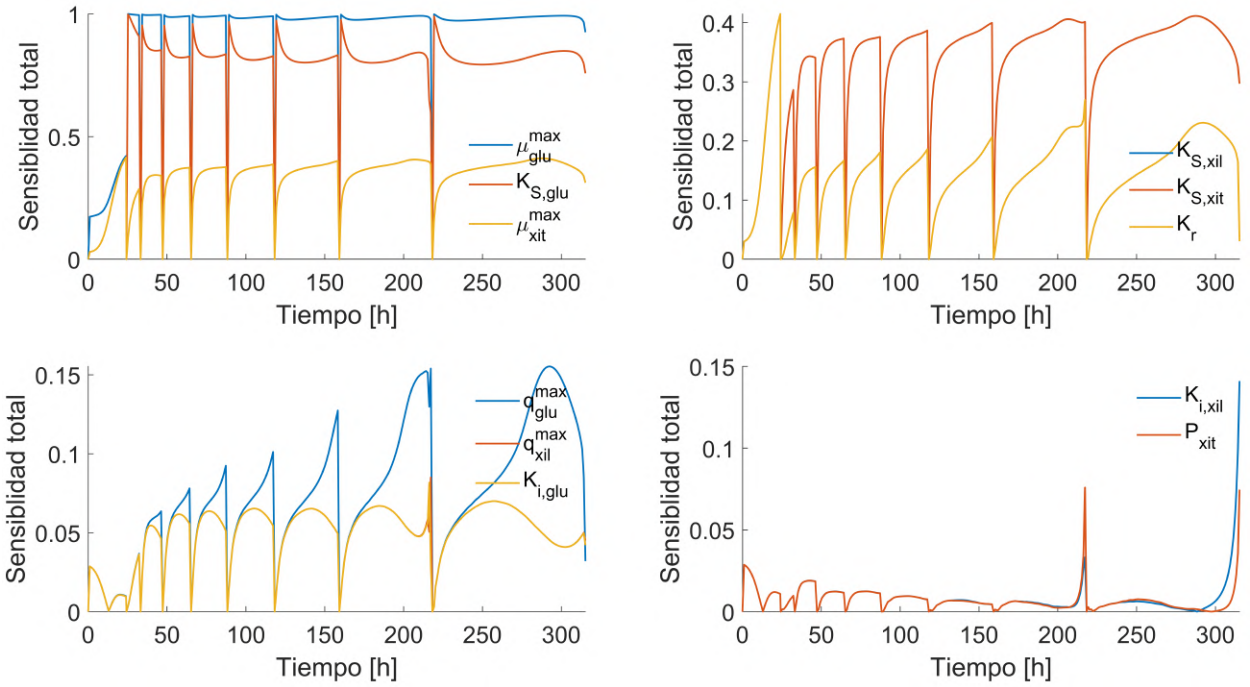


Figura 5-11: Perfiles de índices de sensibilidad totales de Sobol detallados por parámetro con distribución uniforme para el modelo matemático de fermentación diaúxica para producción de xilitol.

El apéndice C.9.2 presenta el análisis de los perfiles de índices de sensibilidad calculados con muestreo realizado con distribución normal y matriz de covarianza. Los perfiles de índices de sensibilidad total presentan valores mayores a 2, lo que indica un doble efecto de la correlación de parámetros. Este resultado surge de la interacción propia de los parámetros sumada

a muestras de parámetros seleccionadas de tal forma que explícitamente presentan la correlación. Debido a lo anterior, no es recomendable el uso de muestreo con distribución normal y matriz de covarianza si los parámetros estimados se encuentran altamente correlacionados.

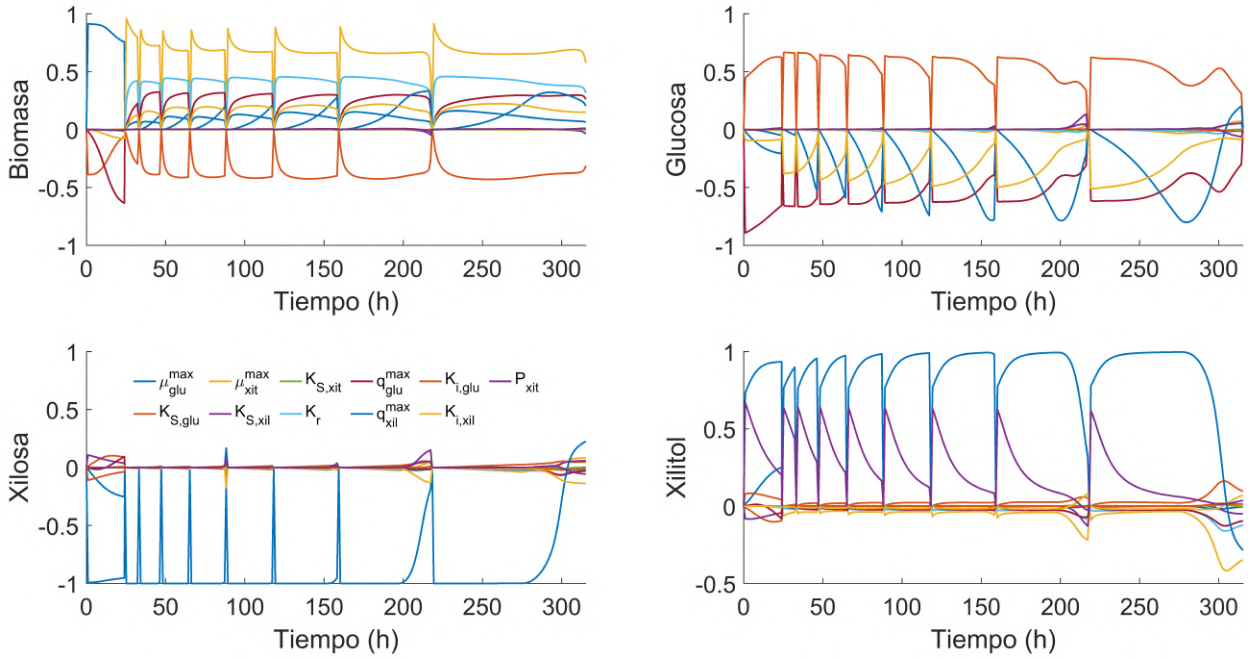


Figura 5-12: Perfiles de índices de sensibilidad SRC para el modelo matemático de fermentación diaúxica para producción de xilitol.

Una forma alternativa para calcular índices de sensibilidad de los parámetros estimados corresponde a los coeficientes de regresión estandarizados (SRC por sus siglas en inglés). La Figura **5-12** presenta los índices SRC para los estados del modelo matemático estudiado. Para estos índices se observan diferencias en la sensibilidad del modelo a diferentes parámetros por estado, situación que no se presentó con los índices de Sobol. Nuevamente se observan diferencias entre el conjunto de datos 1 respecto a los conjuntos de datos 2 a 9, siendo sus parámetros dominantes μ_{glu}^{max} y $K_{S,glu}$ respecto a μ_{xit}^{max}, $K_{S\,glu}$, K_r, q_{glu}^{max}, $K_{i,xil}$, μ_{glu}^{max}. En los conjuntos de datos 2 a 9 los parámetros μ_{gly}^{max}, $K_{S,glu}$ y q_{glu}^{max} presentan influencia, aún cuando la concentración de este sustrato es nula. Lo anterior expone claramente correlación entre los parámetros del modelo, como se indica en la Tabla **C-2**. Para los demás estados del modelo, los parámetros con mayor sensibilidad son $K_{S,glu}$, q_{glu}^{max}, q_{xil}^{max} y $K_{i,xil}$ para glucosa, q_{xil}^{max} para xilosa y finalmente, q_{xil}^{max} y P_{xit} para xilitol. Una vez más, los resultados mostrados indican que en caso de que los parámetros del modelo matemático se encuentren altamente correlacionados, los índices SRC permiten una mejor aproximación a los perfiles de sensibilidad de los parámetros. Sin embargo, dado que los índices SRC son calculados a partir de una regresión lineal múltiple, debe cumplirse un $R^2 \geq 0.7$ para que el índice sea confiable (Storlie & Helton, 2008).

La Figura **5-13** presenta el valor de R^2 correspondiente a los índices SRC de Figura **5-12**. En términos generales, el indicador R^2 presenta un valor constante cercano a 1. Sin embargo, se presentan puntos en donde desciende rápidamente. Específicamente, este comportamiento se da al final de algunos conjuntos de datos, siendo más evidente en los conjuntos de datos 8 y 9 y los estados de xilosa y xilitol. Este fenómeno sería producto de la correlación entre parámetros, la cual genera una salida de mayor carácter no lineal a medida que transcurre el tiempo de simulación.

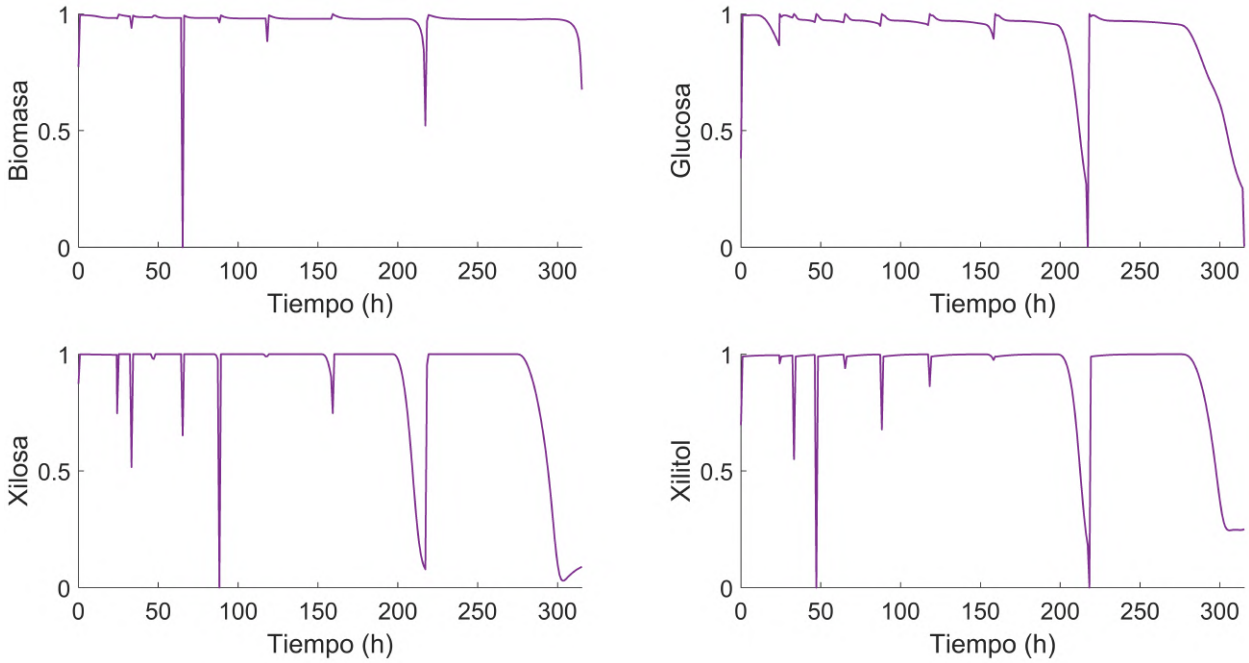


Figura 5-13: Perfil de indicador R^2 de regresión lineal múltiple de índices SRC calculados.

5.4.7. Directrices de estimación y validación

Los modelos matemáticos de procesos biotecnológicos comúnmente pueden incluirse en la categoría de modelos fenomenológicos, los cuales incorporan mecanismos o fenómenos del sistema y cuentan con parámetros cuyo valor debe ser calculado a partir de datos experimentales. En este sentido, la estimación de parámetros corresponde al proceso de encontrar el valor numérico de los parámetros del modelo a partir de un conjunto de datos experimentales. Dado que la interpretabilidad de un parámetro depende tanto de su significado físico como de su valor numérico, menor incertidumbre en la magnitud del parámetro conlleva a incrementar la interpretabilidad numérica de este. La validación de modelos corresponde entonces al cálculo de la incertidumbre que presenta el modelo en sus parámetros y respuesta una vez se encuentra completamente especificado. Tradicionalmente, se ha considerado como parte del proceso de estimación de parámetros la aplicación de métodos de validación

usando los mismos datos experimentales utilizados para la calibración del modelo. De igual forma, se ha considerado como validación de modelos a las mismas metodologías aplicadas a conjuntos de datos diferentes a los utilizados en la calibración del modelo.

Con el objetivo de aportar mayor claridad a la interpretabilidad de los parámetros y a la calidad del modelo, los procesos de estimación de parámetros y validación de modelos serán divididos en estimación de parámetros, verificación de calidad descriptiva y verificación de calidad predictiva. La estimación de parámetros comprende la solución del problema de optimización, es decir, obtener el valor numérico de los parámetros. La verificación de calidad descriptiva implica el cálculo de la incertidumbre de parámetros y respuesta del modelo junto con la comparación de la respuesta del modelo con información ya conocida por el mismo (datos experimentales utilizados en estimación). La verificación de la capacidad predictiva implica la comparación de la respuesta del modelo con información desconocida por el mismo (datos experimentales no utilizados en estimación) y el cálculo de la sensibilidad del modelo a incertidumbre en sus parámetros, esto es lo que comúnmente se denomina validación. Con base en la Figura **5-1**, las directrices de estimación de parámetros y validación de modelos se presentan a continuación:

1. Estimación de parámetros: este proceso involucra los diferentes elementos del problema de optimización, a saber, datos experimentales, función objetivo, algoritmo de optimización global y modelo matemático. Por tanto, las directrices descritas en las secciones 2.4.4. y 4.4.3. también son aplicables.

    1.1. Realizar pretratamiento de datos experimentales (ver sección 2.4.4.).

    1.2. Realizar análisis de identificabilidad estructural. En caso de que el modelo no sea estructuralmente identificable es necesario modificar el diseño experimental o reparametrizar el modelo (si es posible) y volver al paso 1.1.

    1.3. Definir función objetivo (estimador).

    1.4. Definir espacio paramétrico Θ. El rango de los parámetros puede establecerse mediante información previa, metodologías como la plateada por Pitt & Banga (2019) o definiendo un espacio lo suficientemente amplio. Una mejor selección del espacio paramétrico puede aumentar la eficiencia de búsqueda y reducir interacción entre parámetros.

    1.5. Selección de datos experimentales para estimación. Lenard Ljung sugiere utilizar un 30 % del total de datos para estimación de parámetros y el 70 % restante para validación del modelo, repartidos de forma aleatoria (Ljung, 1990). Sin embargo, la selección de la cantidad de datos experimentales depende de la naturaleza del experimento (fermentación batch, semicontinua, continua). Dado lo anterior, se recomienda iniciar la estimación de parámetros con el 30 % de los experimentos

y aumentar su cantidad si no se logra identificabilidad práctica (paso 2.1). Adicionalmente, si el experimento es de tipo batch la selección de los datos puede ser aleatoria (experimentos independientes). En caso de que la fermentación sea de tipo semicontinua o continua, se recomienda seleccionar los datos en orden cronológico.

1.6. Realizar sintonización de algoritmo de optimización (ver sección 4.4.3.).

1.7. Resolver el problema de optimización para estimación de parámetros.

Una vez se han estimado los parámetros, el modelo matemático se encuentra completamente especificado y es posible analizar su comportamiento.

2. Verificación de calidad descriptiva: un modelo matemático fenomenológico o semi-físico se construye como una hipótesis del funcionamiento del sistema modelado. Por otra parte, dado que el modelo ha incorporado información del sistema (biológico y/o físico) a través sus parámetros, idealmente este debería estar en la capacidad de replicar o "describir" el comportamiento del sistema. La verificación de la calidad descriptiva se emplea para corroborar o refutar el modelo como hipótesis del funcionamiento del sistema, es decir, que el sistema efectivamente funciona bajo los consideraciones descritas en el modelo matemático al tener la capacidad de describir información que ya conoce.

   2.1. Calcular intervalos de confianza para los parámetros (Fórmula 5-18). En caso de que el parámetro presente un intervalo de confianza que incluya cero, el parámetro se clasifica como prácticamente no identificable e indica insuficiencia de información experimental. Es posible aplicar metodologías de diseño óptimo de experimentos (DOE) para reducir la incertidumbre en los parámetros. Un parámetro con un nivel de incertidumbre bajo (e.g. inferior al 10 %) se considera interpretable. Para el caso de modelos matemáticos basados en ecuaciones diferenciales ordinarias, se requiere de experimentos con nuevas condiciones iniciales. Posteriormente, se regresa al paso 1.1.

   2.2. Calcular matriz de correlación (Fórmula 5-19). Valores cercanos a —1— indican alta interacción entre parámetros. Existen metodologías para reducir estas interacciones como la re-estimación de parámetros con función objetivo basadas en la matriz de covarianza (García et al., 2017).

   2.3. Calcular indicadores de ajuste. El coeficiente de determinación (R^2) lineal puede ser utilizado como una medida rápida de la capacidad descriptiva del modelo, en donde valores que tienden a 1 indican una adecuada capacidad descriptiva. Adicionalmente, otros indicadores pueden ser utilizados para calcular la discrepancia general del modelo respecto a los datos experimentales. Bajo homogeneidad relativa en número de datos y su escala entre experimentos es recomendable utilizar

los indicadores SSE, RMSE y FPE. En caso de que haya diferencias en número de datos o escalas de magnitud se recomiendan los indicadores MRE y GoF.

2.4. Calcular intervalos de confianza del predictor (Fórmula 5-29). Entre más estrecho sea el intervalo de incertidumbre, mejor predicción en la respuesta del modelo. En caso de que los datos experimentales se encuentren dentro del intervalo de confianza del predictor y su tendencia coincida con un valor de R^2 no tan cercano a 1, se podría suponer que existen inconvenientes con los datos experimentales. En caso de que la tendencia de la respuesta no coincida con los datos experimentales, se presentaría error por parte del modelo. Si se presentan problemas con los datos experimentales regresar al paso 1.1. Si por el contrario el modelo no coincide con los datos experimentales se requiere replantear el modelo.

2.5. Calcular análisis de autocorrelación (Fórmula 5-30). Este análisis permite determinar si el modelo ha fallado en capturar toda la información contenida en los datos experimentales y es complementario al paso 2.4. En caso de que la prueba de autocorrelación falle es necesario verificar la calidad de los datos experimentales o replantear el modelo matemático.

2.6. Calcular índices de sensibilidad. Los perfiles de importancia paramétrica (Fórmula 5-33) permiten establecer la relevancia de un parámetro de forma dinámica para un determinado estado. Los perfiles son válidos si el $R^2 \geq 0.7$. Adicionalmente, si los parámetros no están altamente correlacionados es posible usar los índices de Sobol (Fórmulas 5-31 y 5-32) para determinar la sensibilidad general del modelo respecto a los parámetros.

Si el modelo matemático cumple satisfactoriamente con los pasos 2.1 a 2.6 se puede afirmar que cuenta con capacidad descriptiva, por tanto, se corrobora la hipótesis de que el sistema funciona bajo el mecanismo descrito por el modelo matemático. Además, los parámetros del modelo serían interpretables al poseer un significado físico (dado en la formulación del modelo), ser estructuralmente identificables (su valor puede ser conocido) y ser prácticamente identificables (baja incertidumbre).

3. Verificación de calidad predictiva: Una vez se ha validado que el modelo presenta capacidad descriptiva, es necesario determinar la calidad predictiva del mismo. Si el modelo matemático presenta capacidad predictiva puede ser utilizado en aplicaciones prácticas tales como diseño, optimización y control. Sin embargo, el modelo tendrá aplicabilidad siempre y cuando el sistema no se modifique (i.e. sea invariable en el tiempo).

3.1. Calcular indicadores de ajuste. Se realiza con datos experimentales diferentes a los utilizados en la calibración del modelo (estimación de parámetros). En caso de que el valor de los índices (SSE, RMSE, FPE, MRE o GoF) calculados sea similar al obtenido en el paso 2.3, sería indicativo de que el modelo presenta algún grado de capacidad predictiva. No obstante, es necesario seguir al paso 3.2.

3.2. Calcular intervalos de confianza del predictor (Fórmula 5-29). Si el comportamiento de la respuesta del modelo sigue la tendencia de los datos experimentales y presenta una desviación consistente, sería indicativo de que el sistema se ha modificado. En este caso el sistema es variante en el tiempo y el valor de los parámetros se ha modificado. Se debe regresar al paso 1.1 utilizando parte de los datos experimentales de validación para re-estimación de parámetros. En caso de que la respuesta del modelo no siga la tenencia de los datos experimentales, indicaría la posible aparición de nuevos fenómenos en el sistema, por tanto, el modelo ya no sería aplicable y carecería de capacidad predictiva.

3.3. Calcular análisis de autocorrelación (Fórmula 5-30). Un análisis de autocorrelación con fallos solo alrededor del retraso cero, podría indicar que la estructura del modelo es correcta, pero que el sistema se ha modificado (e.g. ha cambiado el valor de los parámetros). Este análisis es complementario al paso 3.2.

5.5. Conclusiones

La estimación de parámetros o identificabilidad práctica de un modelo matemático comprende varios elementos como lo es modelo matemático, función objetivo, algoritmo de optimización, solucionador y datos experimentales que en conjunto constituyen el problema de optimización para estimación de parámetros. Dada la cantidad de componentes que constituyen el problema de optimización y su influencia sobre el valor estimado de los parámetros, una aproximación sistemática es requerida.

En el presente trabajo se analizó la interconexión entre los diferentes elementos y su influencia sobre el valor estimado de los parámetros del modelo matemático de fermentación diaúxica de glucosa y xilosa para producción de xilitol. Se determinó que la identificabilidad estructural del modelo es una condición necesaria pero no suficiente para alcanzar identificabilidad práctica. En este sentido, la incertidumbre de los parámetros y su influencia en la capacidad descriptiva y predictiva del modelo matemático estudiado fue calculada a través de diferentes pruebas estadísticas. Se determinó que información experimental insuficiente en calidad o cantidad puede generar alta incertidumbre en los parámetros, así mismo, los intervalos de confianza del predictor pueden proporcionar una medida adecuada de la capacidad descriptiva y predictiva del modelo. En caso de que las trayectorias del modelo matemático y los datos experimentales coincidan dentro de los intervalos de confianza, el modelo presenta alta calidad.

De manera específica y con base en intervalos de confianza, los parámetros estimados presentan alta capacidad descriptiva, sin embargo, el modelo carece de capacidad predictiva en mayor grado para la concentración de xilosa y en menor grado para la concentración de

xilitol. Dicha situación pudo relacionarse con el diseño del experimento, demostrando la relevancia y necesidad de un diseño experimental apropiado basado en el modelo matemático que pretende ser construido.

Así mismo, para el caso de estudio se encontró una alta correlación entre parámetros, dada por la naturaleza no lineal del modelo y carencia de información experimental, específicamente fermentaciones con glucosa. Se determinó adicionalmente, que el muestreo con distribución uniforme por muestreo por hipercubo latino (LHS por sus siglas en inglés) es eficiente para el cálculo de índices de sensibilidad de Sobol con estimador "Oracle" de Monte Carlo. No obstante, para el modelo analizado se encontró que los índices de sensibilidad de tipo coeficientes de regresión estandarizados (SRC por sus siglas en inglés) permiten un mayor entendimiento del comportamiento de los parámetros estimados y su influencia en el modelo matemático, bajo alta correlación de parámetros.

Finalmente, se propone la metodología de estimación de parámetros de modelos de procesos biotecnológicos presentada en la Figura **5-1** la cual asegura la consecución de un modelo matemático con interpretabilidad y capacidad tanto descriptiva como predictiva.

Bibliografía

Abt, V., Barz, T., Cruz-Bournazou, M. N., Herwig, C., Kroll, P., Möller, J., Pörtner, R., & Schenkendorf, R. (2018). Model-based tools for optimal experiments in bioprocess engineering. Current opinion in chemical engineering, 22, 244–252.

Aster, R. C., Borchers, B., & Thurber, C. H. (2005). Parameter estimation and inverse problems. Elsevier.

Barrigón, J. M., Ramon, R., Rocha, I., Valero, F., Ferreira, E. C., & Montesinos, J. L. (2012). State and specific growth estimation in heterologous protein production by pichia pastoris. AIChE journal, 58(10), 2966–2979.

Biegler, L. T. (2010). Nonlinear programming: concepts, algorithms, and applications to chemical processes, volume 10. Siam.

Bogaerts, P. & Wouwer, A. (2004). Parameter identification for state estimation—application to bioprocess software sensors. Chemical Engineering Science, 59(12), 2465–2476.

Braniff, N., Scott, M., & Ingalls, B. (2019). Component characterization in a growth-dependent physiological context: optimal experimental design. Processes, 7(1), 52.

Cameron, I. T. & Hangos, K., Eds. (2001). Process Modelling and Model Analysis, volume 4. Academic press.

Dayarian, A., Chaves, M., Sontag, E. D., & Sengupta, A. M. (2009). Shape, size, and robustness: feasible regions in the parameter space of biochemical networks. PLoS Comput Biol, 5(1), e1000256.

DiStefano III, J. (2015). Dynamic systems biology modeling and simulation. Academic Press.

Dochain, D. (2013). Automatic control of bioprocesses. John Wiley & Sons.

Dorantes-Landa, D. N., Cocotle-Ronzón, Y., Morales-Cabrera, M. A., & Hernández-Martínez, E. (2020). Modeling of the xylitol production from sugarcane bagasse by immobilized cells. Journal of Chemical Technology & Biotechnology, 95(7), 1936–1945.

Englezos, P. & Kalogerakis, N. (2000). Applied parameter estimation for chemical engineers. CRC Press.

García, M. R., Alonso, A. A., & Balsa-Canto, E. (2017). A normalisation strategy to optimally design experiments in computational biology. In International Conference on Practical Applications of Computational Biology & Bioinformatics (pp. 126–136).: Springer.

Gelman, A., Goodrich, B., Gabry, J., & Vehtari, A. (2019). R-squared for bayesian regression models. The American Statistician, 73(3), 307–309.

González-Figueredo, C., Flores-Estrella, R. A., & Rojas-Rejón, O. A. (2018). Fermentation: Metabolism, kinetic models, and bioprocessing. In Current topics in biochemical engineering. IntechOpen.

Good, P. I. & Hardin, J. W. (2012). Common errors in statistics (and how to avoid them). John Wiley & Sons.

Gunay, A. (2007). Application of nonlinear regression analysis for ammonium exchange by natural (bigadiç) clinoptilolite. Journal of Hazardous Materials, 148(3), 708–713.

Hasenauer, J., Breindl, C., Waldherr, S., & Allgöwer, F. (2010). Approximative classification of regions in parameter spaces of nonlinear odes yielding different qualitative behavior. In 49th IEEE Conference on Decision and Control (CDC) (pp. 4114–4119).: IEEE.

Hassan, S. S., Farhan, M., Mangayil, R., Huttunen, H., & Aho, T. (2013). Bioprocess data mining using regularized regression and random forests. BMC systems biology, 7(1), 1–7.

Holst, J., Holst, U., Madsen, H., & Melgaard, H. (1993). Validation of grey box models. In Adaptive Systems in Control and Signal Processing 1992 (pp. 53–60). Elsevier.

Johnson, M. L. (2010). Essential numerical computer methods. Academic Press.

Katz, L., Chen, Y. Y., Gonzalez, R., Peterson, T. C., Zhao, H., & Baltz, R. H. (2018). Synthetic biology advances and applications in the biotechnology industry: a perspective. Journal of industrial microbiology & biotechnology, 45(7), 449–461.

Khan, I., Hou, F., & Le, H. P. (2020). The impact of natural resources, energy consumption, and population growth on environmental quality: Fresh evidence from the united states of america. Science of The Total Environment, 754, 142222.

Koutinas, M., Kiparissides, A., Pistikopoulos, E. N., & Mantalaris, A. (2012). Bioprocess systems engineering: transferring traditional process engineering principles to industrial biotechnology. Computational and structural biotechnology journal, 3(4), e201210022.

Kreutz, C., Raue, A., Kaschek, D., & Timmer, J. (2013). Profile likelihood in systems biology. The FEBS journal, 280(11), 2564–2571.

Kroll, P., Hofer, A., Stelzer, I. V., & Herwig, C. (2017). Workflow to set up substantial target-oriented mechanistic process models in bioprocess engineering. Process Biochemistry, 62, 24–36.

Kucherenko, S., Feil, B., Shah, N., & Mauntz, W. (2011). The identification of model effective dimensions using global sensitivity analysis. Reliability Engineering & System Safety, 96(4), 440–449.

Kucherenko, S. & Song, S. (2017). Different numerical estimators for main effect global sensitivity indices. Reliability Engineering & System Safety, 165, 222–238.

Kucherenko, S., Tarantola, S., & Annoni, P. (2012). Estimation of global sensitivity indices for models with dependent variables. Computer physics communications, 183(4), 937–946.

Ljung, L. (1990). System Identification: Theory for the User. Prentice-Hall.

Matlab (2018). Global optimization toolbox. Mathworks.

MatLab (2018). isoutlier() command documentation. MathWorks.

Mohamad, N. L., Kamal, S. M. M., Mokhtar, M. N., Husain, S. A., & Abdullah, N. (2016). Dynamic mathematical modelling of reaction kinetics for xylitol fermentation using candida tropicalis. Biochemical Engineering Journal, 111, 10–17.

Morales-Rodriguez, R., Meyer, A. S., Gernaey, K. V., & Sin, G. (2012). A framework for model-based optimization of bioprocesses under uncertainty: Lignocellulosic ethanol production case. Computers & Chemical Engineering, 42, 115–129.

Moser, A., Kuchemüller, K. B., Deppe, S., Rodríguez, T. H., Frahm, B., Pörtner, R., Hass, V. C., & Möller, J. (2021). Model-assisted doe software: optimization of growth and biocatalysis in saccharomyces cerevisiae bioprocesses. Bioprocess and biosystems engineering, 44(4), 683–700.

Myers, R. H., Montgomery, D. C., Vining, G. G., & Robinson, T. J. (2012). Generalized linear models: with applications in engineering and the sciences, volume 791. John Wiley & Sons.

Nielsen, F., Tomás-Pejó, E., Olsson, L., & Wallberg, O. (2015). Short-term adaptation during propagation improves the performance of xylose-fermenting saccharomyces cerevisiae in simultaneous saccharification and co-fermentation. Biotechnology for Biofuels, 8(1), 219.

Nocedal, J. & Wright, S. (2006). Numerical optimization. Springer Science & Business Media.

O'Brien, C. M., Zhang, Q., Daoutidis, P., & Hu, W.-S. (2021). A hybrid mechanistic-empirical model for in silico mammalian cell bioprocess simulation. Metabolic Engineering, 66, 31–40.

Pelletier, G. J., Chapra, S. C., & Tao, H. (2006). Qual2kw: A framework for modeling water quality in streams and rivers using a genetic algorithm for calibration. Environmental Modelling & Software, 21(3), 419–425.

Pitt, J. A. & Banga, J. R. (2019). Parameter estimation in models of biological oscillators: an automated regularised estimation approach. BMC bioinformatics, 20(1), 82.

Postawa, K., Szczygieł, J., & Kułażyński, M. (2020). A comprehensive comparison of ode solvers for biochemical problems. Renewable Energy.

Prado-Rubio, O. A., Hernández-Escoto, H., Rodriguez-Gomez, D., Sirisansaneeyakul, S., & Morales-Rodriguez, R. (2015). Enhancing xylitol bio-production by an optimal feeding policy during fed-batch operation. In Computer Aided Chemical Engineering, volume 37 (pp. 1757–1762). Elsevier.

Qian, G. & Mahdi, A. (2020). Sensitivity analysis methods in the biomedical sciences. Mathematical Biosciences, 323, 108306.

Sadino-Riquelme, M. C., Rivas, J., Jeison, D., Hayes, R. E., & Donoso-Bravo, A. (2020). Making sense of parameter estimation and model simulation in bioprocesses. Biotechnology and Bioengineering, 117(5), 1357–1366.

Sainz, J., Pizarro, F., Pérez-Correa, J. R., & Agosin, E. (2003). Modeling of yeast metabolism and process dynamics in batch fermentation. Biotechnology and Bioengineering, 81(7), 818–828.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & Tarantola, S. (2010). Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. Computer physics communications, 181(2), 259–270.

Savitzky, A. & Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. Analytical chemistry, 36(8), 1627–1639.

Schuler, M. M. & Marison, I. W. (2012). Real-time monitoring and control of microbial bioprocesses with focus on the specific growth rate: current state and perspectives. Applied microbiology and biotechnology, 94(6), 1469–1482.

Shampine, L. F. & Reichelt, M. W. (1997). The matlab ode suite. SIAM journal on scientific computing, 18(1), 1–22.

Shieh, Y.-Y. & Fouladi, R. T. (2003). The effect of multicollinearity on multilevel modeling parameter estimates and standard errors. Educational and psychological measurement, 63(6), 951–985.

Sirisansaneeyakul, S., Wannawilai, S., & Chisti, Y. (2013). Repeated fed-batch production of xylitol by candida magnoliae tistr 5663. Journal of Chemical Technology & Biotechnology, 88(6), 1121–1129.

Sobol, I. (1993). Sensitivity estimates for nonlinear mathematical models. Mathematical Modelling and Computational Experiments, 1(4), 407–414.

Sobol, M. & Myshetskaya, E. E. (2008). Monte carlo estimators for small sensitivity indices. Monte Carlo Methods and Applications, 13(5-6), 455–465.

Soize, C. (2017). Uncertainty quantification. Springer.

Storlie, C. B. & Helton, J. C. (2008). Multiple predictor smoothing methods for sensitivity analysis: Description of techniques. Reliability Engineering & System Safety, 93(1), 28–54.

Sun, N.-Z. & Sun, A. (2015). Model calibration and parameter estimation: for environmental and water resource systems. Springer.

Telen, D., Logist, F., Van Derlinden, E., Tack, I., & Van Impe, J. (2012). Optimal experiment design for dynamic bioprocesses: a multi-objective approach. Chemical Engineering Science, 78, 82–97.

Tochampa, W., Sirisansaneeyakul, S., Vanichsriratana, W., Srinophakun, P., Bakker, H. H., Wannawilai, S., & Chisti, Y. (2015). Optimal control of feeding in fed-batch production of xylitol. Industrial & Engineering Chemistry Research, 54(7), 1992–2000.

Urniezius, R. & Survyla, A. (2019). Identification of functional bioprocess model for recombinant e. coli cultivation process. Entropy, 21(12), 1221.

Van den Bos, A. (2007). Parameter estimation for scientists and engineers. John Wiley & Sons.

Villaverde, A. F. (2019). Observability and structural identifiability of nonlinear biological systems. Complexity, 2019.

Villaverde, A. F., Evans, N. D., Chappell, M. J., & Banga, J. R. (2019). Input-dependent structural identifiability of nonlinear systems. IEEE Control Systems Letters, 3(2), 272–277.

Villez, K., Del Giudice, D., Neumann, M. B., & Rieckermann, J. (2020). Accounting for erroneous model structures in biokinetic process models. Reliability Engineering & System Safety, 203, 107075.

Wang, H. & Wang, X.-c. (2014). Parameter estimation for metabolic networks with two stage bregman regularization homotopy inversion algorithm. Journal of Theoretical Biology, 343, 199–207.

Wei, P., Lu, Z., & Song, J. (2015). Variable importance analysis: a comprehensive review. Reliability Engineering & System Safety, 142, 399–432.

Weise, K. (1985). Uncertainty of parameter estimation. IFAC Proceedings Volumes, 18(5), 1717–1722.

Capítulo 6

Conclusiones y perspectivas

La biotecnología es una herramienta que puede contribuir a la solución de problemas sociales, ambientales y económicos a los que la humanidad ya se está enfrentando, tales como el cambio climático, seguridad alimentaria, escasez de recursos hídricos, producción de medicamentos, entre otros. Los sistemas biológicos pueden ser representados como modelos matemáticos, que desde el punto de vista de ingeniera de sistemas de proceso, pueden ser útiles para el diseño, optimización y control de bioprocesos. Sin embargo, dichos modelos matemáticos son generalmente de tipo no lineal y aunque suelen ser formulados a partir de bases teóricas, requieren de estimación de parámetros. Lo anterior presenta un reto desde el punto de vista matemático, pues para que el modelo sea apto se requiere que sus parámetros sean interpretables y su predicción sea acorde al comportamiento real del sistema biológico.

Para dar respuesta a estas interrogantes, en esta tesis se complementan metodologías existentes y se implementan herramientas para estimación de parámetros y validación de modelos matemáticos de procesos biotecnológicos. Esto con el objetivo de incrementar la interpretabilidad de los parámetros y asegurar las capacidades descriptivas y predictivas de dichos modelos matemáticos. La metodología consolidada analiza cada uno de los elementos que componen el problema de estimación de parámetros de manera independiente y su impacto en los parámetros y la predicción del modelo, como se muestra a continuación:

- Se plateó un procedimiento para la limpieza de datos experimentales de origen biológico, que suelen caracterizarse por presentar un elevado error experimental. Este proceso incluye la detección de puntos atípicos por métodos estadísticos, junto con su reemplazo y la eliminación de ruido aleatorio por medio de métodos de filtrado de señales. Adicionalmente, se propuso la métrica MAPE como indicador de calidad de los datos experimentales "crudos". Es este sentido, los datos experimentales analizados presentaron una diferencia cualitativa en su calidad entre variables, siendo aquellas con mayor calidad las concentraciones de xilosa y xilitol, seguidas de glucosa y finalmente, biomasa biomasa (MAPE promedio de 26.94 %, 13.81 %, 9.11 % y 44.62 %, respectivamente). Lo anterior se atribuyó a la diferencia en los métodos de análisis para metabolitos (cromatografía) y biomasa (gra-

vimetría). Dado que la trayectoria descrita por el modelo matemático ha de ajustarse a los datos experimentales, mejor calidad en estos implica una mejor predicción del modelo.

- Se realizó la recopilación de métodos para análisis identificabilidad estructural de modelos matemáticos dinámicos y se propusieron directrices para el uso de diferentes software según las características del modelo analizado (ecuaciones, estados, parámetros y estados observados). Se determinó que el modelo de caso de estudio (bioproducción de xilitol) es estructuralmente localmente identificable en experimentos con consumo simultáneo de glucosa y xilosa. Sin embargo, solo cinco de los once parámetros son estructuralmente identificables cuando se utiliza solamente xilosa como sustrato. Lo anterior expone la necesidad de un esfuerzo sincrónico e interdependiente en la construcción de modelos matemáticos y diseño experimental.

- Se investigó la influencia tanto del factor de normalización en la función objetivo como de la configuración del optimizador en la precisión y reproducibilidad de la solución del problema de optimización. Para esto una nueva herramienta por interconexión de Matlab® y R® fue propuesta y probada con resultados satisfactorios (mayor precisión y reproducibilidad en el valor de la función objetivo). Específicamente, nueve combinaciones de algoritmo de optimización y normalización de función objetivo fueron sintonizadas, además, se analizó el efecto de los valores o categorías de hiperparámetros en el desempeño del optimizador. También, se comparó el desempeño de optimizadores con configuraciones sintonizadas y por defecto. Para el caso de estudio se determinó que la combinación de optimizador enjambre de partículas sintonizado con función objetivo de mínimos cuadrados y normalización de valor medio presenta los mejores resultados. Adicionalmente, se propuso una directriz para la preselección de algoritmos de optimización y su sintonización en Matlab® para problemas de optimización específicos.

- Se analizó el efecto de los diferentes elementos que componen el problema de optimización sobre el valor de los parámetros y la predicción del modelo de bioproducción de xilitol. Se encontró que efectivamente el pretratamiento de datos experimentales afecta el valor de los parámetros (reduciendo incertidumbre en un 43.2 %) y que el optimizador sintonizado puede encontrar consistentemente la región del óptimo global (compensando su naturaleza estocástica). Los intervalos de confianza de parámetros y salida del modelo junto con indicadores de ajuste presentaron una alta capacidad descriptiva, por ejemplo, valores de coeficiente de determinación R^2 promedio de 0.85, 0.90 y 0.97 para xilosa, xilitol y glucosa, respectivamente. Sin embargo, no se presenta capacidad predictiva para este caso de estudio, lo que indica un cambio en el sistema biológico que se atribuye a adaptación del microorganismo a concentraciones alta del sustrato xilosa. Se determinó que los índices de sensibilidad de Sobol no son adecuados cuando el modelo matemático presenta una alta correlación en sus parámetros, debido a la generación de múltiples interacciones que afectan especialmente a los índices individuales. Sin embargo, se encontró

que el método de coeficientes de regresión estandarizados (SRC por sus siglas en inglés) es conveniente en presencia de alta correlación de parámetros y, además, es útil para analizar la fenomenología subyacente en los modelos matemáticos.

Tomando en consideración lo anterior, el modelo matemático de bioproducción de xilitol demostró cualidades relevantes para esta investigación como alta no linealidad, sobreparametrización, diseño experimental no óptimo y superficie de optimización no convexa. Adicionalmente, se propuso una metodología basada en una serie de directrices para la estimación de parámetros y validación de modelos matemáticos de procesos biotecnológicos, que surgen de la combinación de métodos ya establecidos. La metodología propuesta puede mejorar la calidad e interpretabilidad de los parámetros del modelo matemático a través de la reducción en la incertidumbre de su valor numérico. Dado que los modelos matemáticos hacen parte de las bases de la ingeniería de sistemas de proceso, una mejora en su calidad y capacidad tanta descriptiva como predictiva se verá reflejada en aplicaciones más precisas y robustas. Esto es especialmente relevante en el área de los procesos biotecnológicos, debido a los diferentes retos que presenta la aplicación práctica de sistemas biológicos como alta complejidad, sensibilidad y variabilidad biológica.

De forma general, las directrices generadas en esta investigación corresponden a:

1. Limpieza de datos experimentales:

    - Detección y reemplazo puntos atípicos, que reduce desviaciones de la trayectoria principal de los datos.
    - Eliminación de ruido aleatorio y restricción a las trayectorias del modelo.

2. Identificabilidad estructural:

    - Robustece la interpretabilidad de los parámetros del modelo matemático.
    - Permite reparametrizar el modelo (en ocasiones), eliminando parámetros no identificables.

3. Sintonización de algoritmos de optimización:

    - Selección de optimizadores eficientes para un problema de optimización en particular.
    - La normalización por valor medio de variable experimental genera mejores resultados en la optimización.
    - Mayor eficiencia computacional y reproducibilidad de resultados.

4. Estimación de parámetros y validación de modelos:

- Calibración efectiva y rápida del modelo matemático con optimizador sintonizado.
- Verificación de la calidad descriptiva y predictiva con medidas de incertidumbre en parámetros y simulaciones, que indica si el modelo presenta estructura, valores de parámetros y datos experimentales adecuados.
- Incrementa la robustez de la aplicabilidad del modelo matemático.

Dada la relevancia de los procesos biotecnológicos y el continuo avance en el entendimiento de los sistemas biotecnológicos, sumado a la necesidad de atacar problemáticas como el cambio climático y la escasez de alimentos, el modelamiento matemático se puede considerar como una herramienta valiosa que puede contribuir a alcanzar procesos industriales sostenibles de una manera factible y eficiente. Considerando lo anterior, el avance en el estudio del modelamiento matemático enfocado a procesos biotecnológicos se hace indispensable. Considerando lo anterior, la investigación realizada en esta tesis extiende la posibilidad de explorar nuevas líneas de investigación aplicadas a modelos matemáticos de procesos biotecnológicos. Entre estas se incluyen:

- Diseño de experimentos guiado por identificabilidad estructural: esta aproximación permitiría aprovechar la naturaleza *a priori* de los análisis de identificabilidad estructural para generar diseños experimentales altamente informativos. De esta manera, sería posible conocer el valor numérico del parámetro y reducir su incertidumbre sin la necesidad de aplicar diseño óptimo de experimentos.

- Mejoras o nuevos algoritmos de optimización estocásticos (metaheurísticas) y métodos de sintonización: diversos mecanismos de búsqueda (algoritmos) pueden ser útiles para diferentes problemas de optimización, lo cual es especialmente relevante en problemas de optimización con superficies no convexas o con alta dimensionalidad. Métodos de sintonización más eficientes permitirían una búsqueda más rápida y precisa en la función objetivo, al incorporar de una mejor manera información del problema de optimización en el optimizador.

- Medidas estadísticas no lineales: las aproximaciones lineales son rápidas y fáciles de implementar, sin embargo, presentan un error que puede variar según el grado de no linealidad y las interacciones dentro del modelo matemático, afectando la robustez del modelo en su aplicación práctica. Tomando en cuenta lo anterior, se hace necesario desarrollar métricas estadísticas no lineales que permitan la realización de análisis más robustos que posiblemente llevarían a una mejor calidad del modelo.

Apéndice A

Tratamiendo de datos experimentales

A continuación, se presentan los resultados del pretratamiento de datos para los conjuntos experimentales diferentes a 1, 9, 10 y 22.

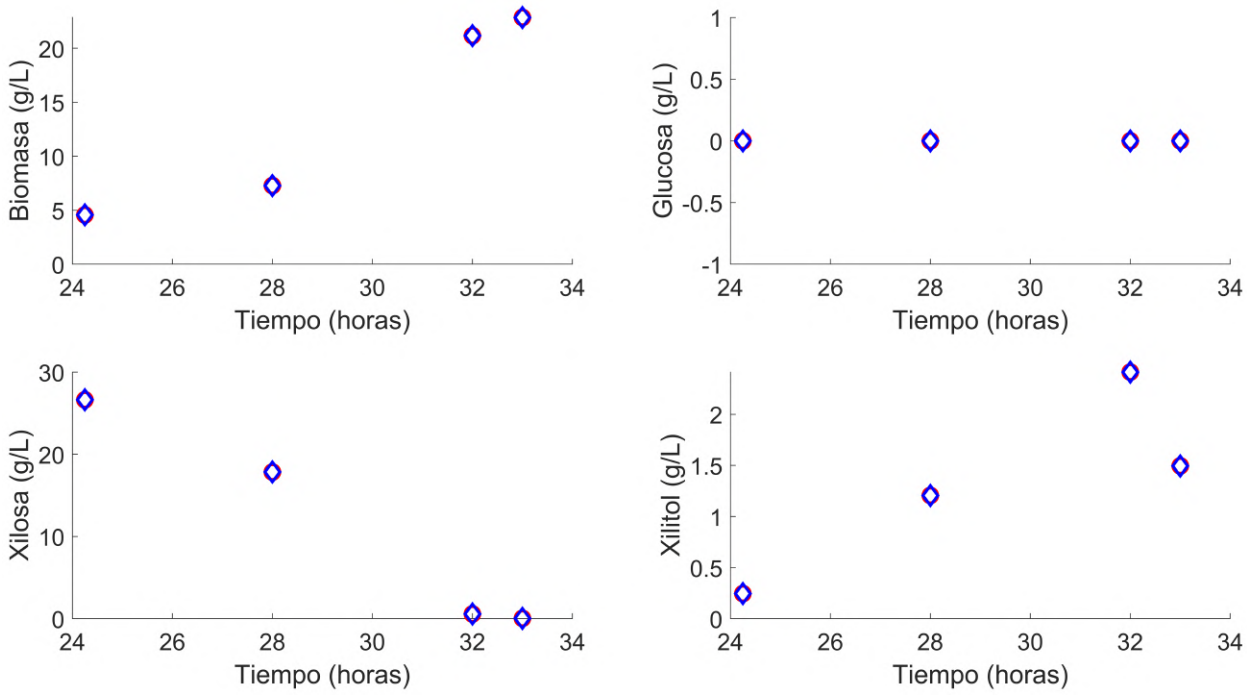


Figura A-1: Conjunto de datos 2: (○) dato experimental, (◇) dato tratado.

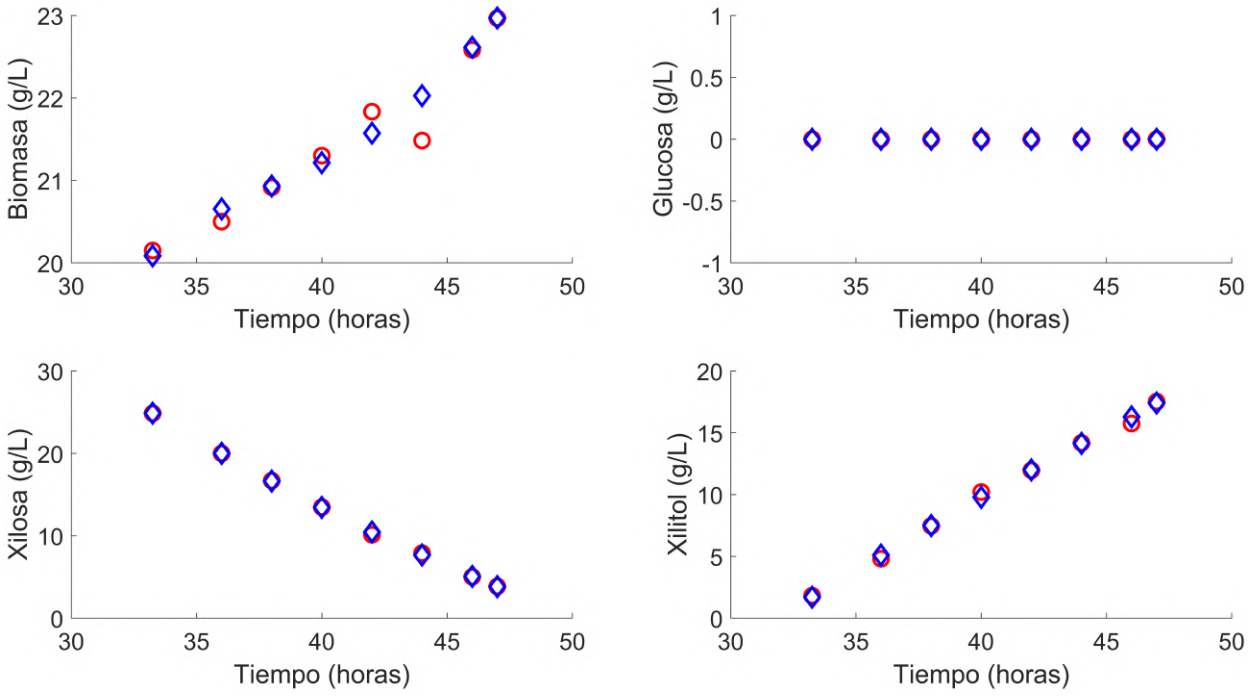


Figura A-2: Conjunto de datos 3: (○) dato experimental, (◇) dato tratado.

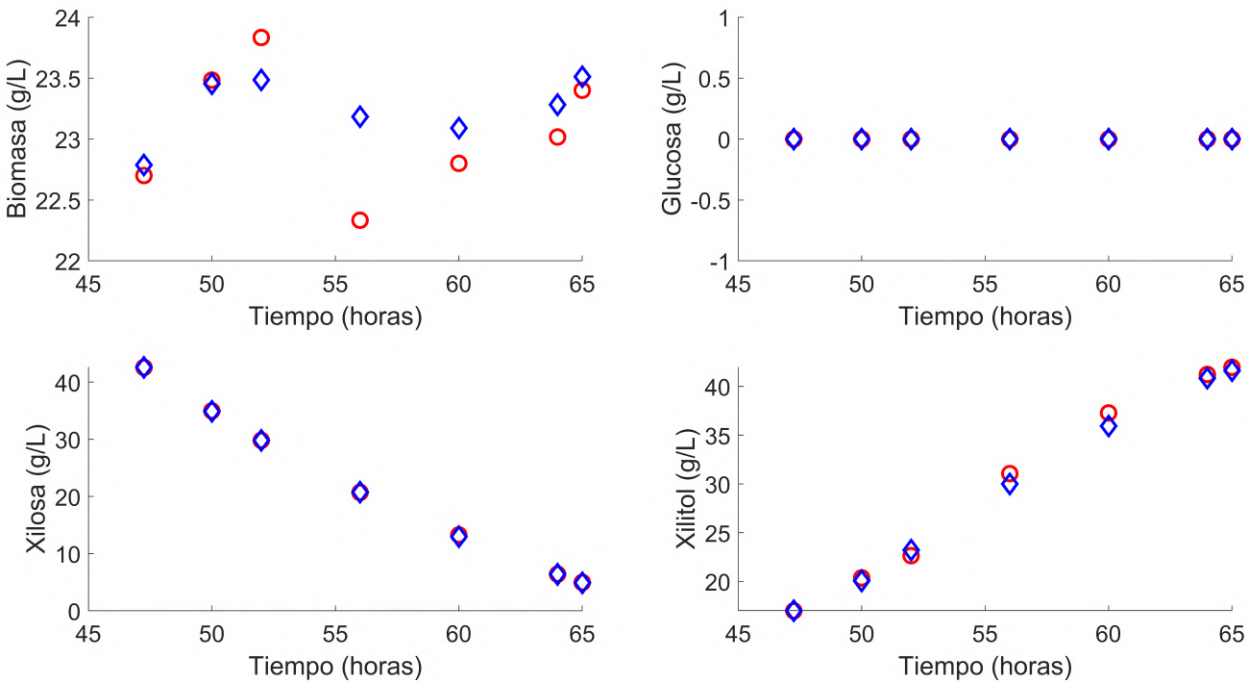


Figura A-3: Conjunto de datos 4: (○) dato experimental, (◇) dato tratado.

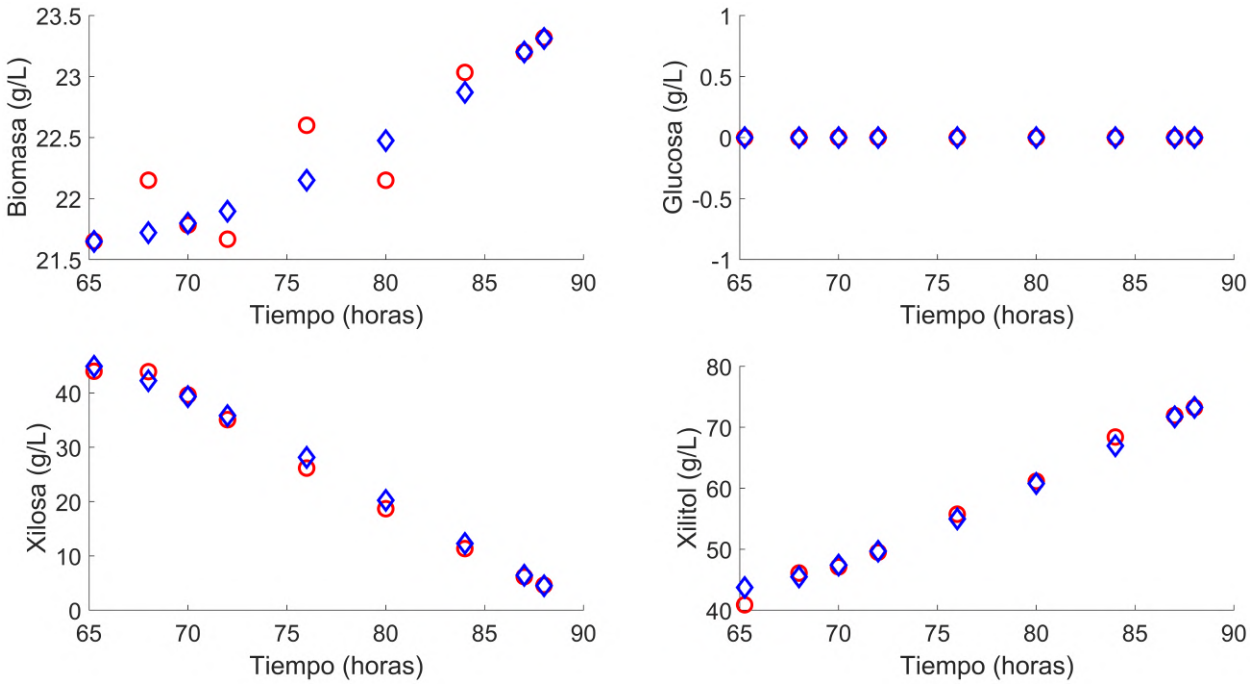


Figura A-4: Conjunto de datos 5: (○) dato experimental, (◇) dato tratado.

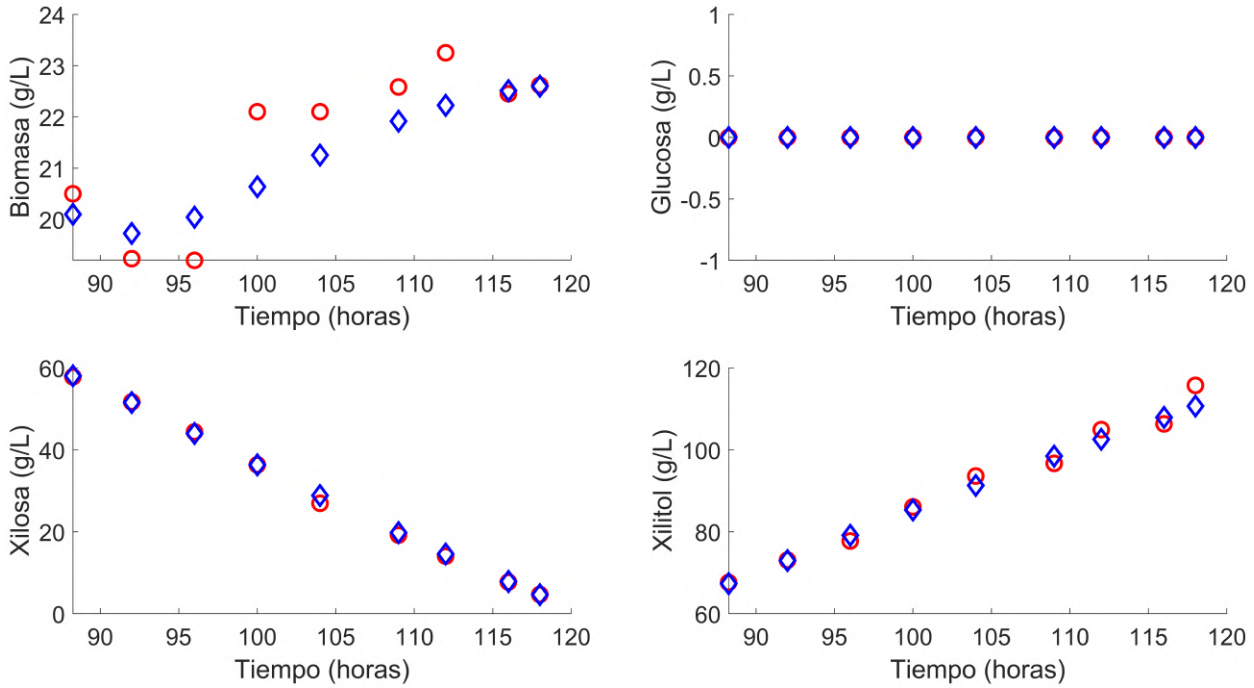


Figura A-5: Conjunto de datos 6: (○) dato experimental, (◇) dato tratado.

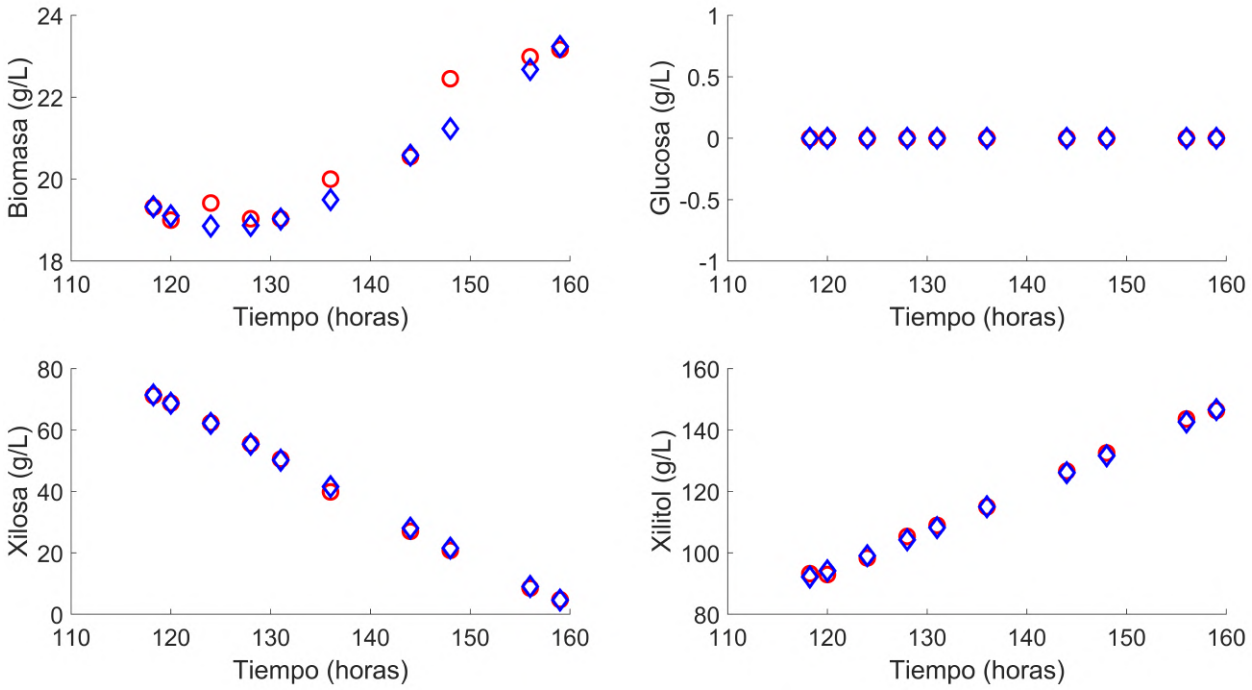


Figura A-6: Conjunto de datos 7: (○) dato experimental, (◇) dato tratado.

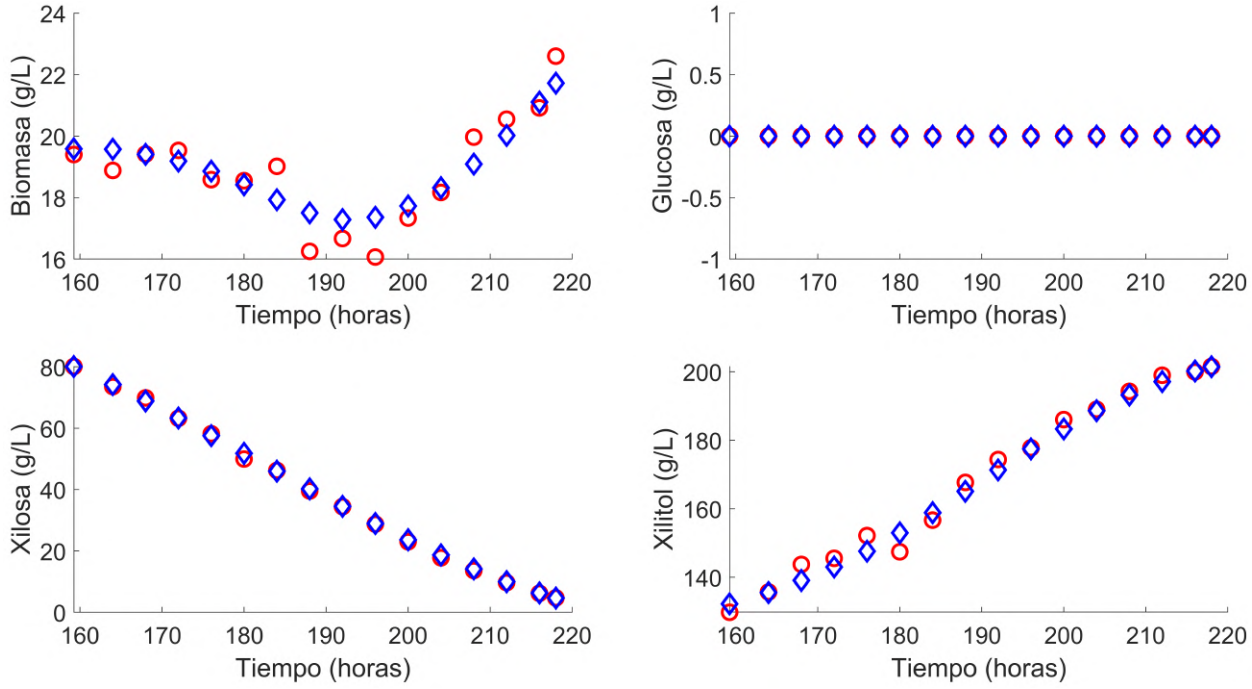


Figura A-7: Conjunto de datos 8: (○) dato experimental, (◇) dato tratado.

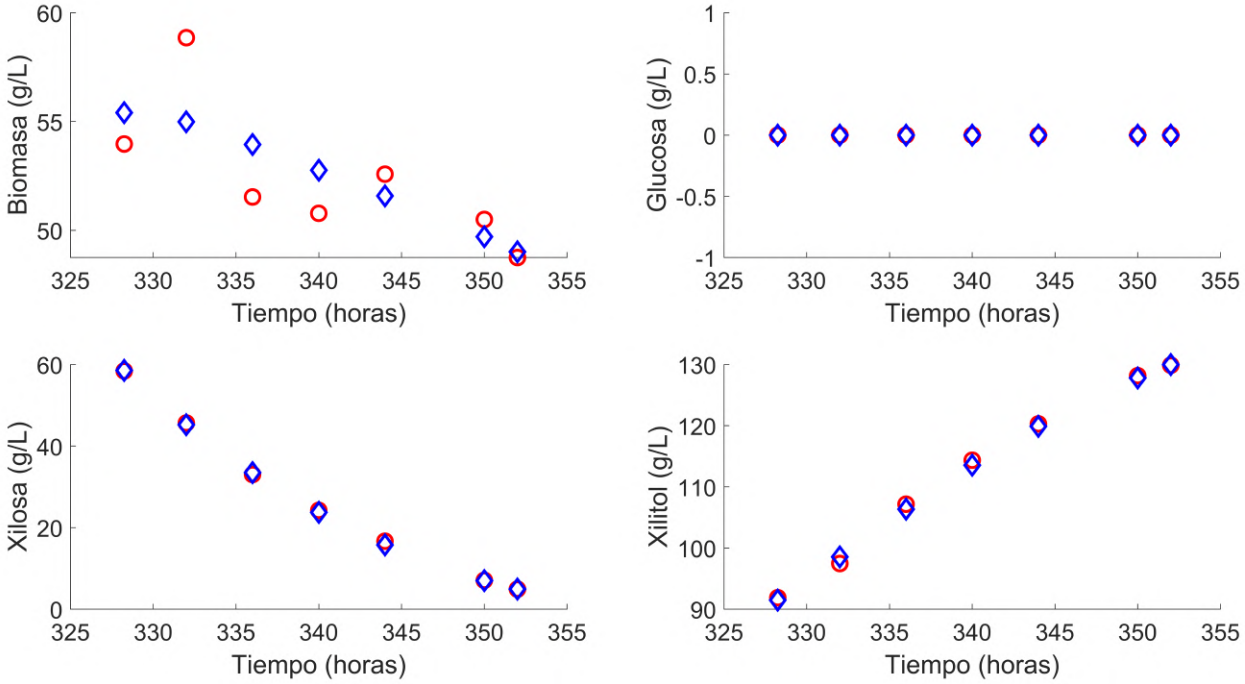


Figura A-8: Conjunto de datos 11: (○) dato experimental, (◇) dato tratado.

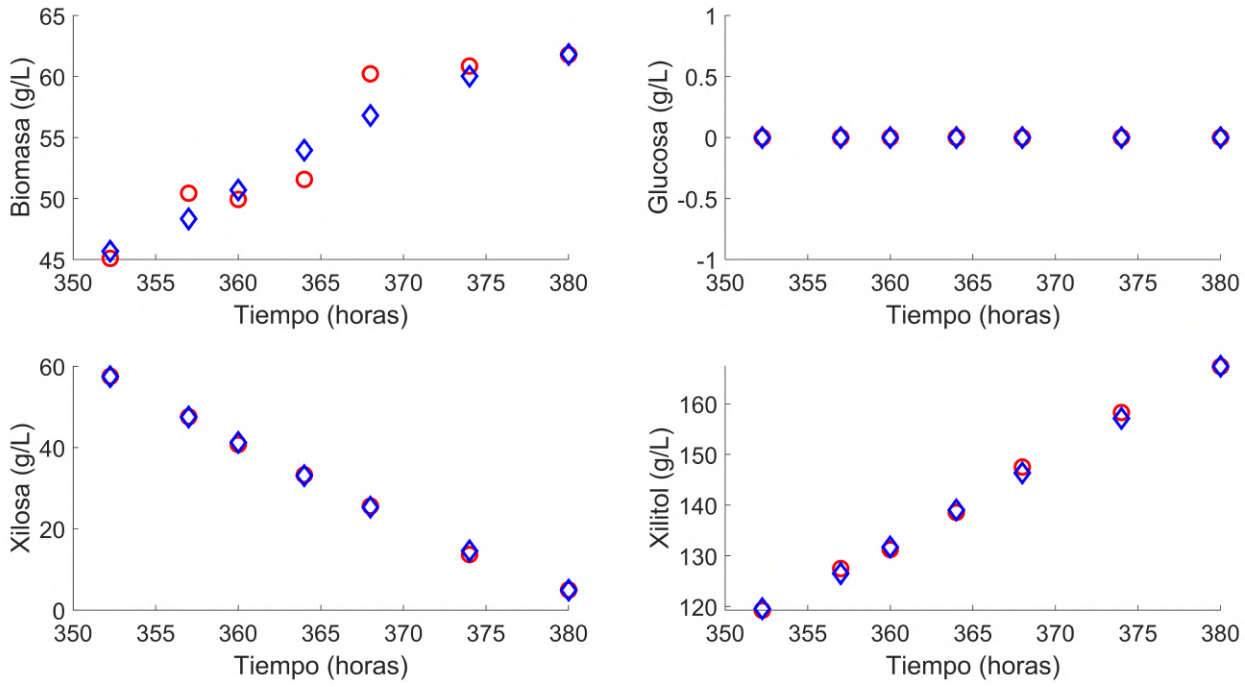


Figura A-9: Conjunto de datos 12: (○) dato experimental, (◇) dato tratado.

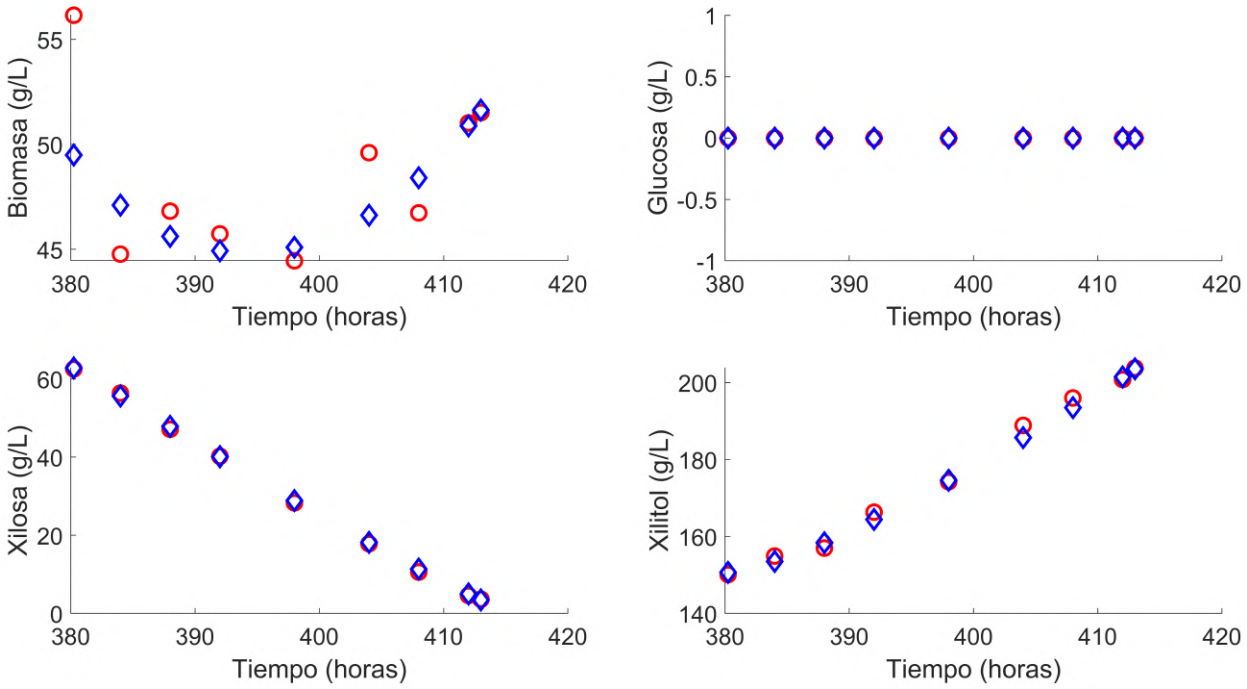


Figura A-10: Conjunto de datos 13: (○) dato experimental, (◇) dato tratado.

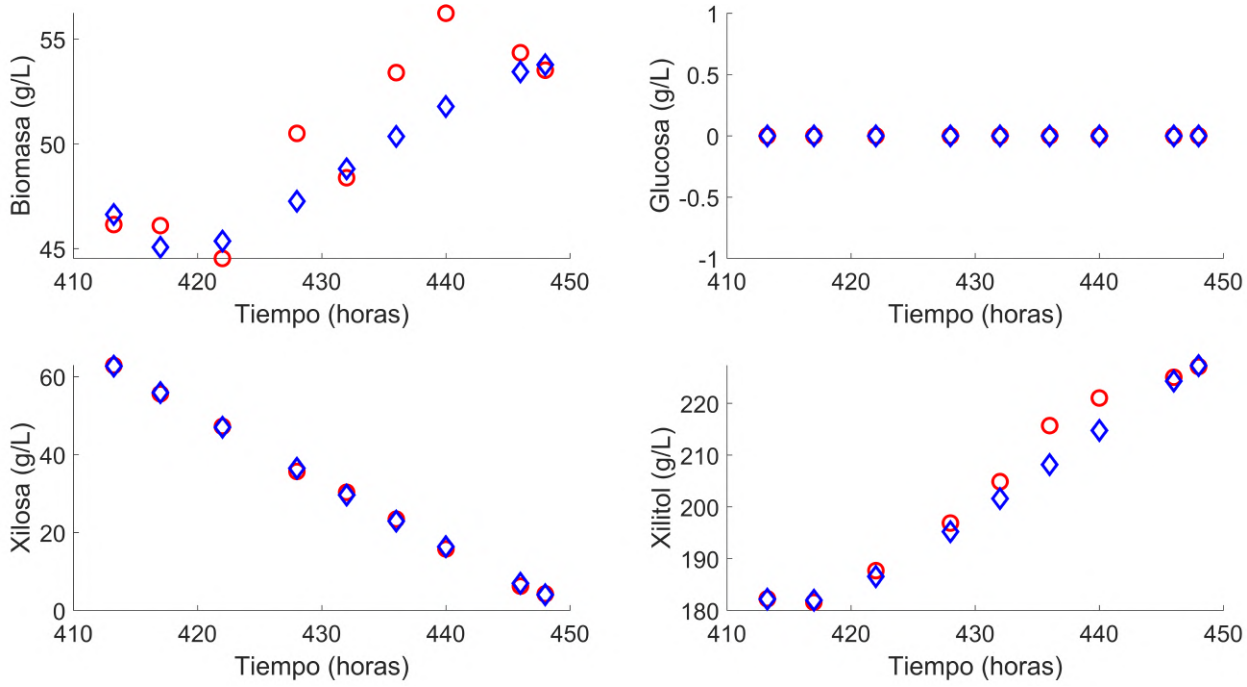


Figura A-11: Conjunto de datos 14: (○) dato experimental, (◇) dato tratado.

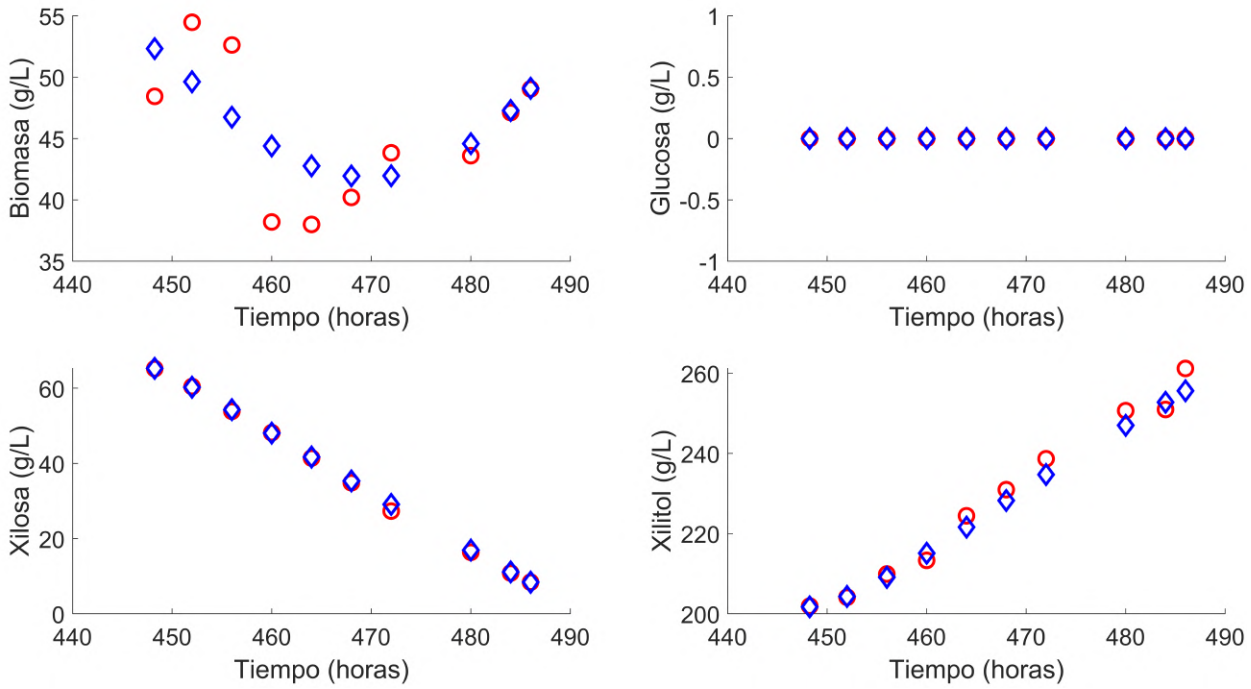


Figura A-12: Conjunto de datos 15: (○) dato experimental, (◇) dato tratado.

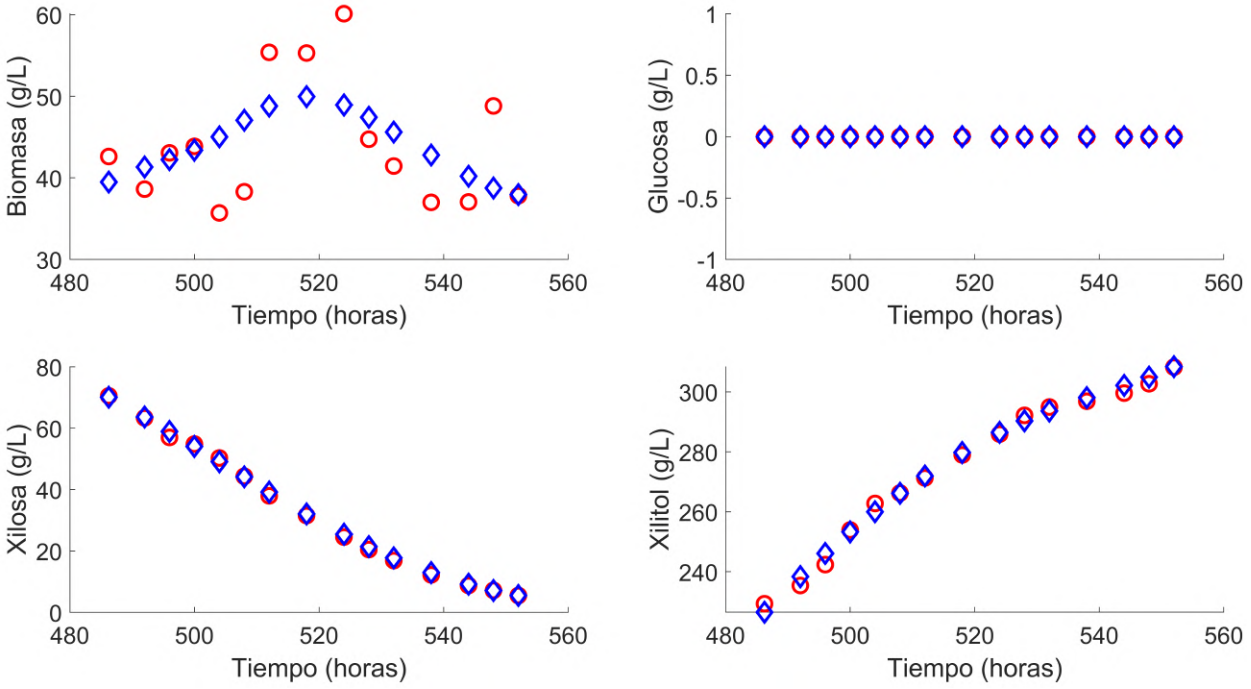


Figura A-13: Conjunto de datos 16: (○) dato experimental, (◇) dato tratado.

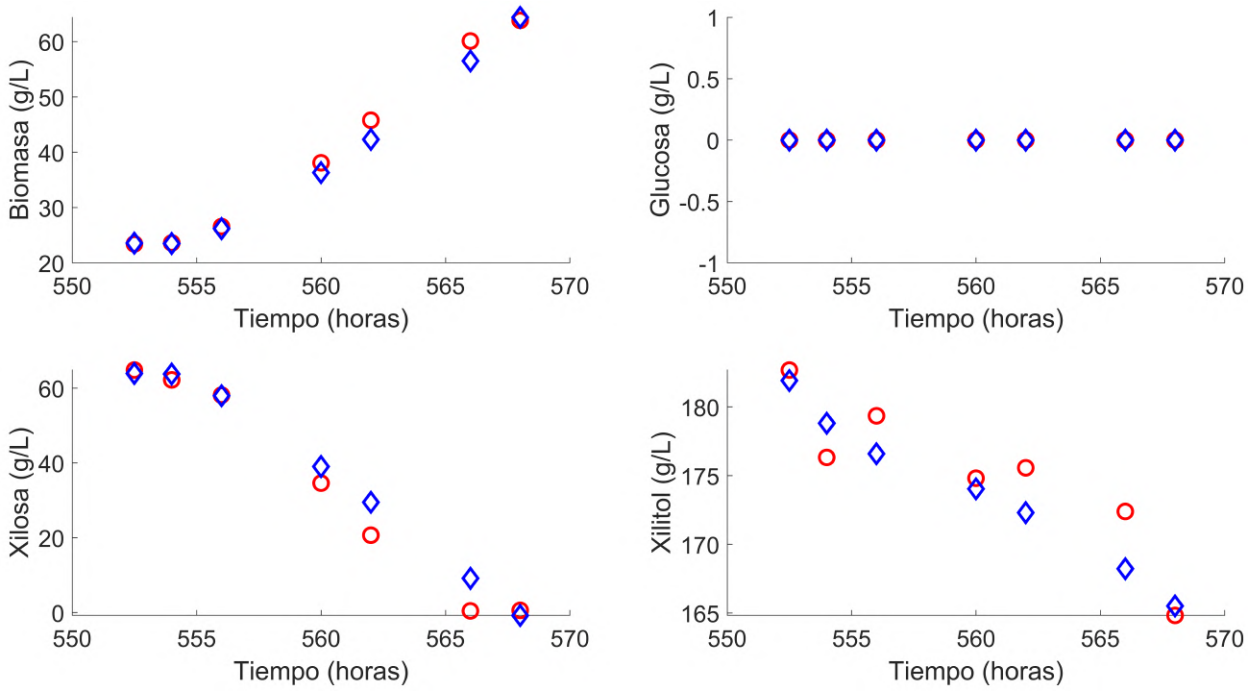


Figura A-14: Conjunto de datos 17: (○) dato experimental, (◇) dato tratado.

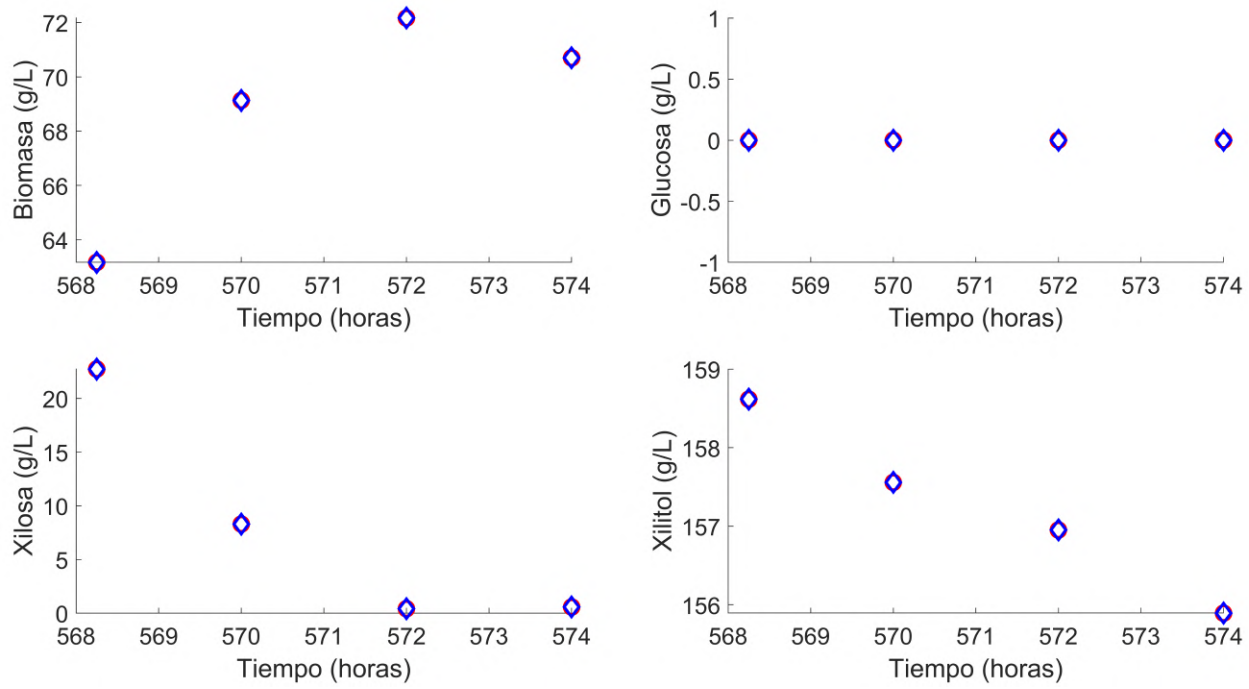


Figura A-15: Conjunto de datos 18: (○) dato experimental, (◇) dato tratado.

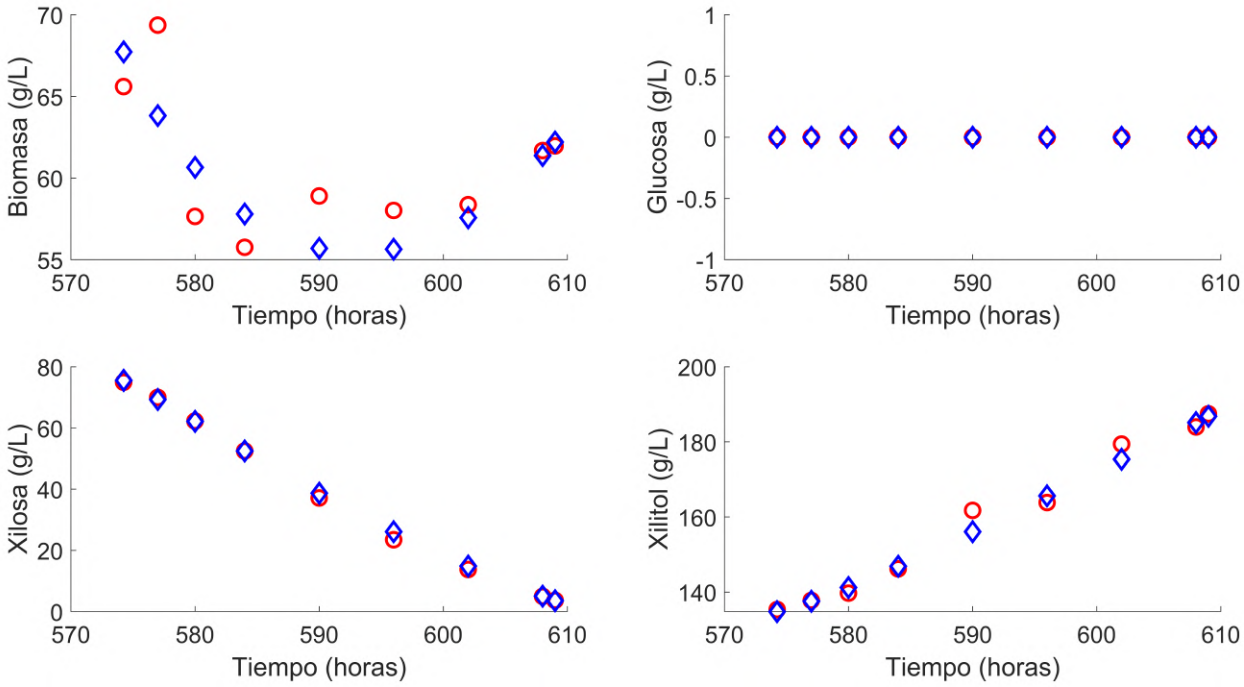


Figura A-16: Conjunto de datos 19: (○) dato experimental, (◇) dato tratado.

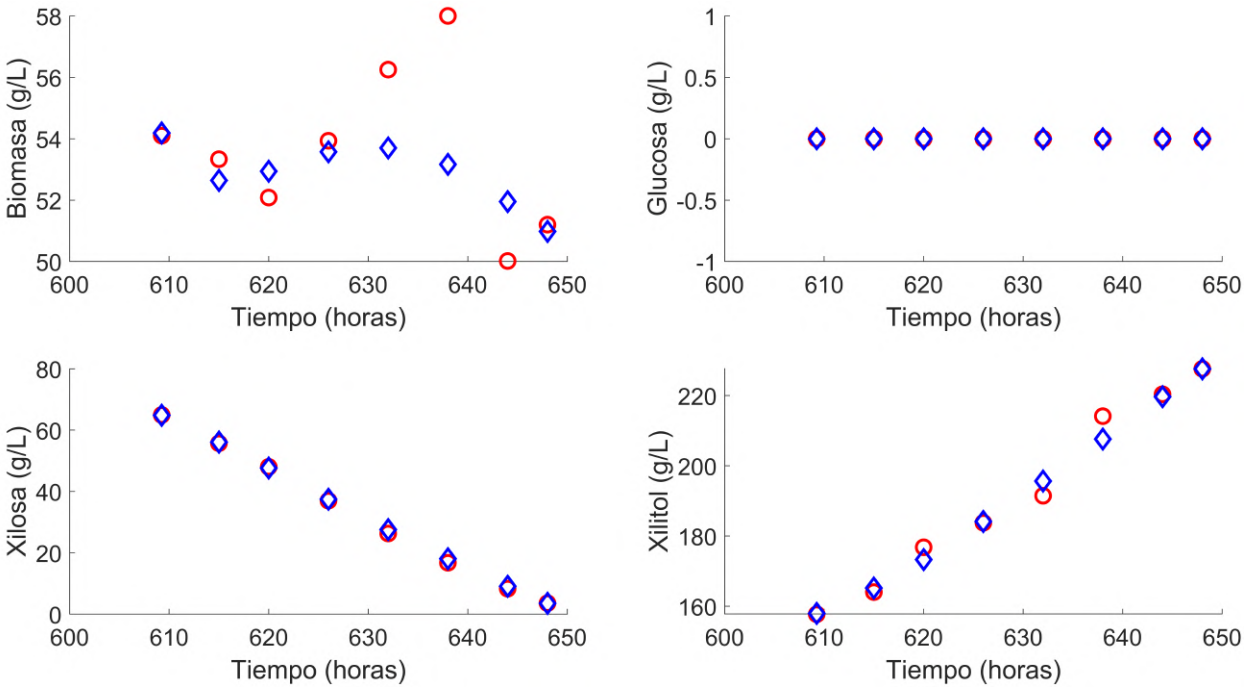


Figura A-17: Conjunto de datos 20: (○) dato experimental, (◇) dato tratado.

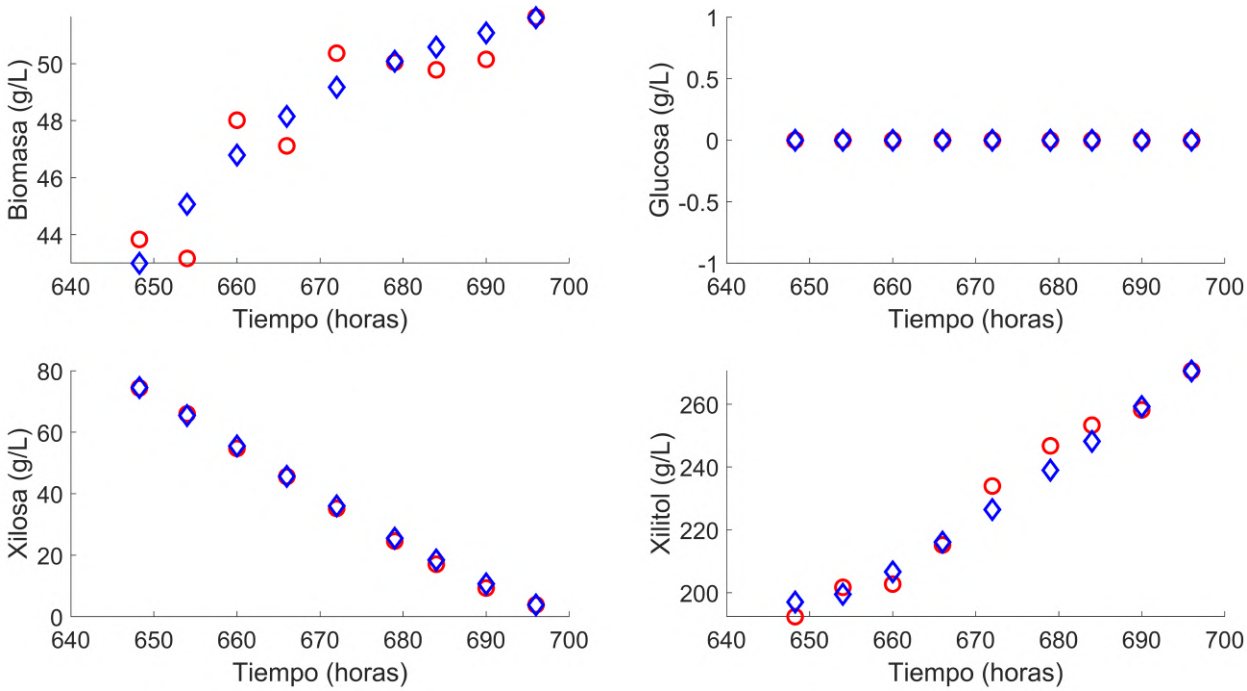


Figura A-18: Conjunto de datos 21: (○) dato experimental, (◇) dato tratado.

Apéndice B

Sintonización de algoritmos de optimización

En el presente apéndice se presentan los resultados completos del análisis de sintonización de algoritmos de optimización.

B.1. Algoritmo detallado de *racing*

Un método que hace parte de la categoría de métodos iterativos es el denominado *racing* (Yuan & Gallagher, 2004). El *racing* es un método de configuración de algoritmos propuesto por Maron y Moore en 1997 y se basa en la evaluación sucesiva de configuraciones posibles para el algoritmo y su respuesta reflejada en la función objetivo (Maron & Moore, 1997). Este proceso hace una evaluación de posibles configuraciones en un total de pruebas y cada configuración evaluada está asociada a la probabilidad de error dada por la frontera de Hoeffding. Maron y Moore proponen además otros criterios diferentes a la frontera de Hoeffding para el proceso de descarte de configuraciones, como el Bayesiano y el Bayesiano bloqueado (Maron & Moore, 1997). Otras modificaciones a este procedimiento básico han sido realizadas, como la inclusión de la prueba no paramétrica de Friedman de dos vías para la comparación entre diferentes configuraciones (denominada *F-Race*) y un proceso de *racing* iterativo (denominado *I-Race*) (Birattari et al., 2002).

A continuación se presentan las ecuaciones para el cálculo de la frontera de Hoeffding como criterio de descarte de configuraciones en el algoritmo *racing*:

$$E_{est,\,conf} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} |\varphi_{conf} - \bar{\varphi}_{conf}| \tag{B-1}$$

$$P(|E_{verd,\,conf} - E_{est,\,conf}| > \varepsilon) < 2e^{-2\,n_{test}\,\varepsilon^2 / B_{conf}^2} \tag{B-2}$$

$$\varepsilon(n_{iter}) = \sqrt{\frac{B_{conf}^2 \log(2\, n_{test}\, n_{conf}) - \log(\Delta)}{2\, n_{test}}} \tag{B-3}$$

$$n_{iter} > \frac{B^2 \log(2/\Delta\, n_{test}\, n_{conf})}{2\, \varepsilon^2} \tag{B-4}$$

$$B = \bar{E}_{est,\, conf} + 2\sigma^2_{E_{est,conf}} \tag{B-5}$$

en donde φ_{conf} indica el valor de función objetivo para una configuración, $\bar{\varphi}_{conf}$ valor medio de la función objetivo para una configuración, $E_{est,\, conf}$ error estimado para una configuración, $E_{verd,\, conf}$ error verdadero para una configuración, ε tolerancia del error de la configuración, $P(|E_{verd,\, conf} - E_{est,\, conf}| > \varepsilon)$ frontera de Hoeffding, n_{test} número de pruebas actual del algoritmo *racing* ($n_{test} < N_{test}$), N_{test} número máximo de pruebas, B error máximo posible de la configuración, n_{conf} número total de configuraciones y Δ probabilidad de que la configuración sea incorrecta en todo el algoritmo.

El algoritmo de *racing* en la iteración i se compone de:

- Selección aleatoria de configuraciones.
- Inicio de iteraciones de racing.
- Prueba del desempeño de las configuraciones φ_{conf}.
- Cálculo del error estimado por configuración $E_{est,\, conf}$.
- Actualización del error estimado promedio por configuración y B
- Calculo de ε y frontera de Hoeffding para el numero de pruebas actual por configuración
- Descarte de las peores configuraciones.
- Finalización del *racing*.

el algoritmo de *racing* puede terminar si se cumple alguna de las siguientes condiciones:

- Todas las configuraciones excepto una han sido eliminadas.
- El número máximo de pruebas N_{test} ha sido alcanzado.
- Las configuraciones remanentes se encuentran en el rango de $2\, \varepsilon\, n_{test}$ (ninguna configuración tiene un mejor desempeño que otra.)

B.2. Funcionamiento básico del algoritmo *irace*

La Figura **B-1** muestra gráficamente el proceso de *racing* implementado en el paquete "irace". En el lado izquierdo I_i serían las pruebas o *instancias* en donde la configuración es evaluada. Los puntos de colores hacen referencias a las configuraciones presentes en dicha instancia. El lado derecho presenta el muestreo de las nuevas configuraciones a partir de distribuciones de probabilidad truncadas. A medida que se evalúan las configuraciones en las

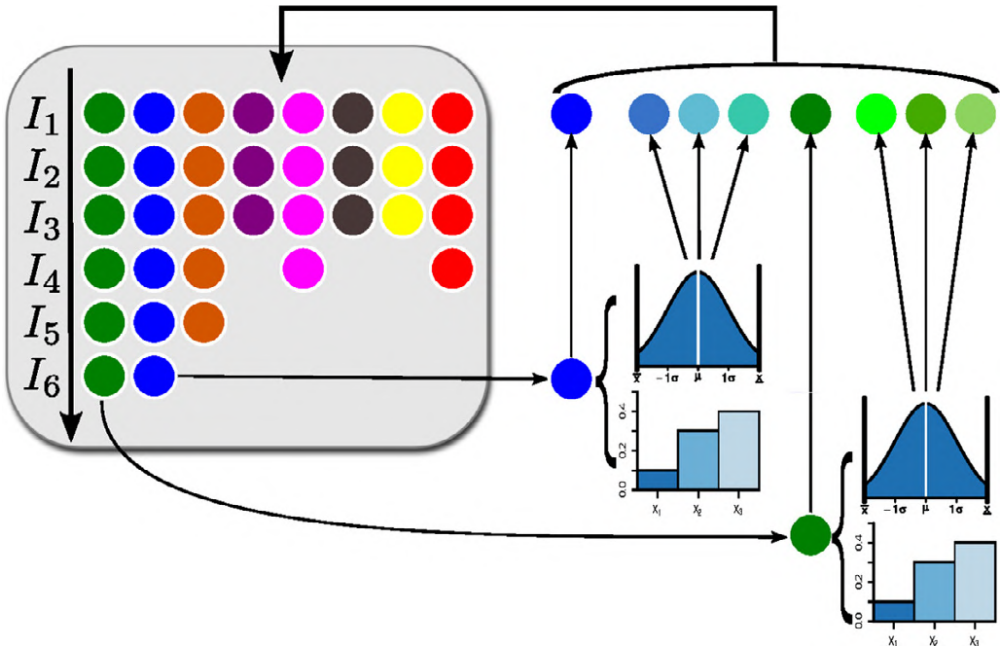
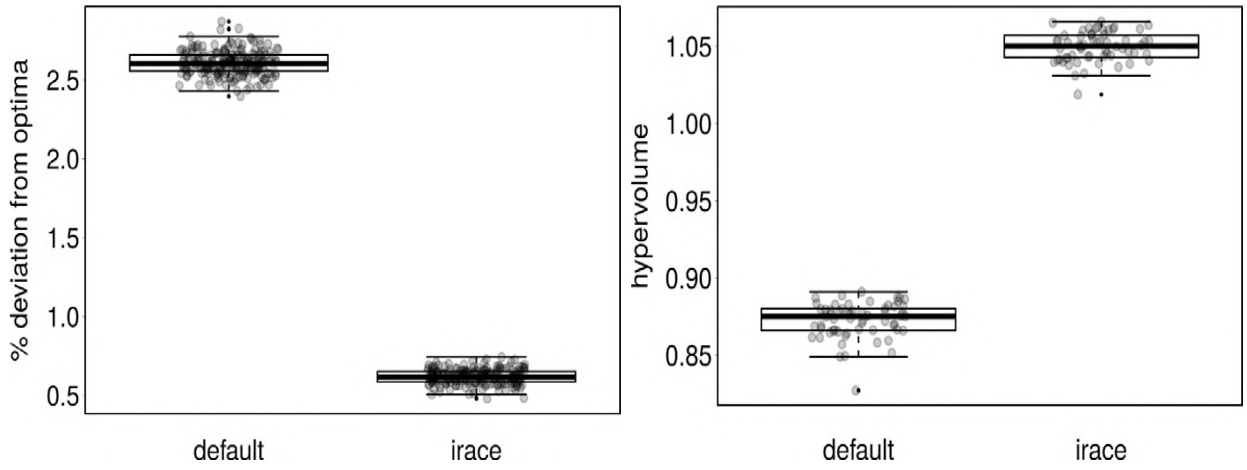


Figura B-1: Esquema gráfico de la metodología *racing* implementada en el paquete "irace", tomada de López-Ibañez, M., Dubois-Lacoste, J., Cáceres, L. P., Birattari, M., & Stützle, T. (2016). The irace package: Iterated racing for automatic algorithm configuration. Operations Research Perspectives, 3, 43-58..



(a) Algoritmo ACOTSP (minimización) (b) Algoritmo MOACO (maximización)

Figura B-2: Ejemplo de resultado obtenido con el paquete "irace", tomado de López-Ibañez, M., Dubois-Lacoste, J., Cáceres, L. P., Birattari, M., & Stützle, T. (2016). The irace package: Iterated racing for automatic algorithm configuration. Operations Research Perspectives, 3, 43-58.

instancias, se realizan pruebas estadísticas para descartar configuraciones como se observa en I_4.

La Figura **B-2** muestra la comparación de la configuración obtenida por "irace" respecto a la configuración por defecto de un optimizador para un problema de minimización (a) y otro de maximización (b). Se observa no solo alta reproducibilidad de los resultados, también una mejora significativa en el valor de la función objetivo.

B.3. Descripción de los optimizadores globales.

B.3.1. Templado simulado

Las técnicas de búsqueda aleatorias corresponden a algoritmos que se basan en cambios aleatorios de las variables de optimización, los cuales se aceptan o descartan según una regla establecida. En esta categoría se encuentra el algoritmo de templado simulado o *simmulated annealing* (SA). Este algoritmo fue propuesto por primera vez en 1989 por Aarts & Korst (1989) y se basa en imitar ciertos principios termodinámicos en la producción de un cristal ideal. Para llevar a cabo el proceso de cristalización la temperatura T debe disminuir. Este parámetro es de vital importancia en este algoritmo pues un cambio rápido de T puede llevar a irregularidades en el cristal, y desde el punto de vista de la optimización, a un resultado fallido. Se requiere entonces de un esquema de enfriamiento que permita un cambio de temperatura adecuado para hallar el mínimo global en un tiempo de cómputo aceptable. El algoritmo básico de SA se muestra en la Figura **B-3**, en donde j es el contador de disminución de temperatura, k contador interno de pasos aleatorios, $F_obj(S^k)$ valor de función objetivo evaluado en S^k. La Tabla **B-1** presenta los parámetros que controlan el comportamiento del algoritmo SA en el *global optmization toolbox* de Matlab® R2018.

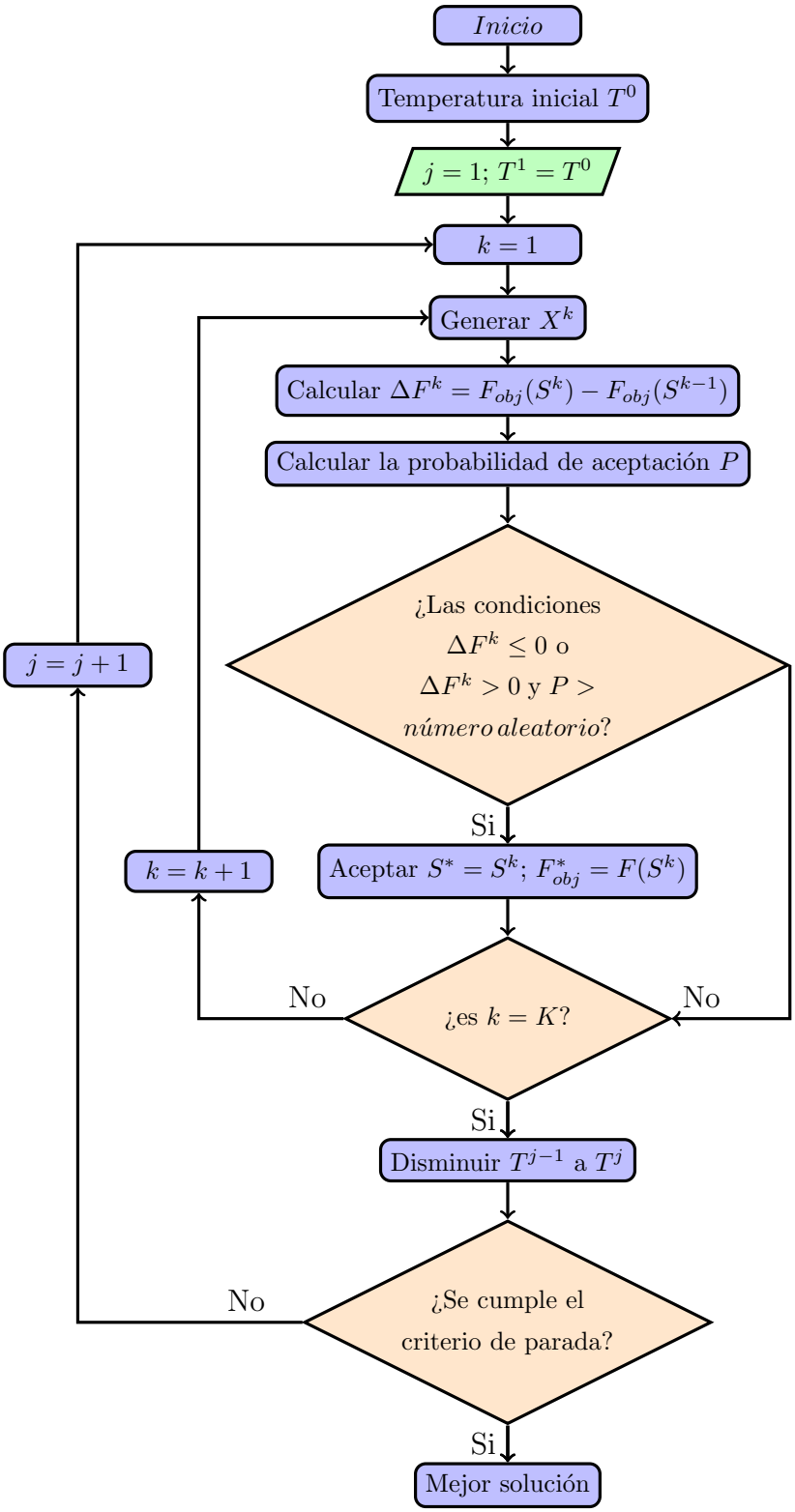


Figura B-3: Algoritmo del optimizador templado simulado (SA).

Tabla B-1: Descripción de los parámetros internos de SA.

`AnnealingFcn`	Controla la generación de nuevos puntos en la próxima iteración.	`annealingfast`	El paso tiene como longitud la temperatura, con dirección uniformemente aleatoria.
		`annealingboltz`	El paso tiene como longitud la raíz cuadrada de la temperatura, con dirección uniformemente aleatoria.
`InitialTemperature`	Temperatura inicial al inicio del algoritmo, controla el grado de búsqueda global (exploración) y búsqueda local (explotación).		
`TemperatureFcn`	Actualiza el esquema de cambio de temperatura. k es el parámetro de templado equivalente al valor de la iteración actual.	`temperatureexp`	La temperatura is igual a `InitialTemperature`*0.95^k
		`temperaturefast`	La temperatura es igual a `InitialTemperature`/k
		`temperatureboltz`	La temperatura es igual a `InitialTemperature`/$ln(k)$
`ReannealInterval`	Número de puntos aceptados antes de realizar nuevamente el templado (*reannealing*).		

B.3.2. Enjambre de partículas

Los métodos de inteligencia de enjambre se basan en replicar el comportamiento social de grupos de seres vivos vistos en la naturaleza como peces, aves e insectos. Los animales gregarios son capaces de resolver problemas de optimización del mundo real, tales como búsqueda de alimento más cercano y asignación de tareas. Este tipo de algoritmos fueron creados a partir de la simulación de los fenómenos sociales como evasión de colisiones, igualación de posiciones y velocidades, unión al centro del grupo, entre otros (Reynolds, 1987).

Entre los algoritmos de inteligencia de enjambre más utilizados se encuentra el algoritmo de optimización por enjambre de partículas o *Particle Swarm Optimization* (PSO), propuesto por Kennedy & Eberhart (1995). Este algoritmo se caracteriza por un conjunto de entidades (partículas) que se mueven de manera aleatoria en el espacio de búsqueda. La velocidad y dirección de las partículas se cambia de manera aleatoria en cada iteración del algoritmo. El comportamiento social de las partículas es introducido mediante fenómenos de unión al centro del enjambre (aproximación a la partícula con mejor desempeño) y evasión de colisiones (búsqueda local realizada por la partícula). Este algoritmo posee grandes ventajas en cuanto a la calidad de la solución y la velocidad de convergencia respecto a otros algoritmos de optimización global (Rangaiah, 2010). El algoritmo de PSO es mostrado en la Figura **B-4**, φ_1 y φ_2 hacen referencia a los términos sociales de unión y repulsión, respectivamente. w es la inercia de las partículas, $\mathbf{r_1}$ y $\mathbf{r_2}$ corresponden a vectores aleatorios, g identificador de mejor partícula global, j contador de comparaciones respecto a mejor partícula global, t contador de iteraciones del optimizador. La Tabla **B-2** presenta los parámetros que controlan el comportamiento del algoritmo PSO en el *global optmization toolbox* de Matlab® R2018.

Tabla B-2: Descripción de los parámetros internos de PSO.

`InertiaRangelb`	Valor inferior de la inercia adaptativa. La inercia es la tendencia de la partícula a permanecer en su posición actual.
`InertiaRangeub`	Valor superior de la inercia adaptatia. La inercia es la tendencia de la partícula a permanecer en su posición actual.
`MinNeighborsFraction`	Tamaño mínimo de vecindario adaptativo.
`SelfAdjustmentWeight`	Término de ponderación de la mejor posición de cada partícula en el ajuste de la velocidad.
`SocialAdjustmentWeight`	Término de ponderación de la mejor posición del vecindario en el ajuste de la velocidad.
`SwarmSize`	Número de partículas en el enjambre.

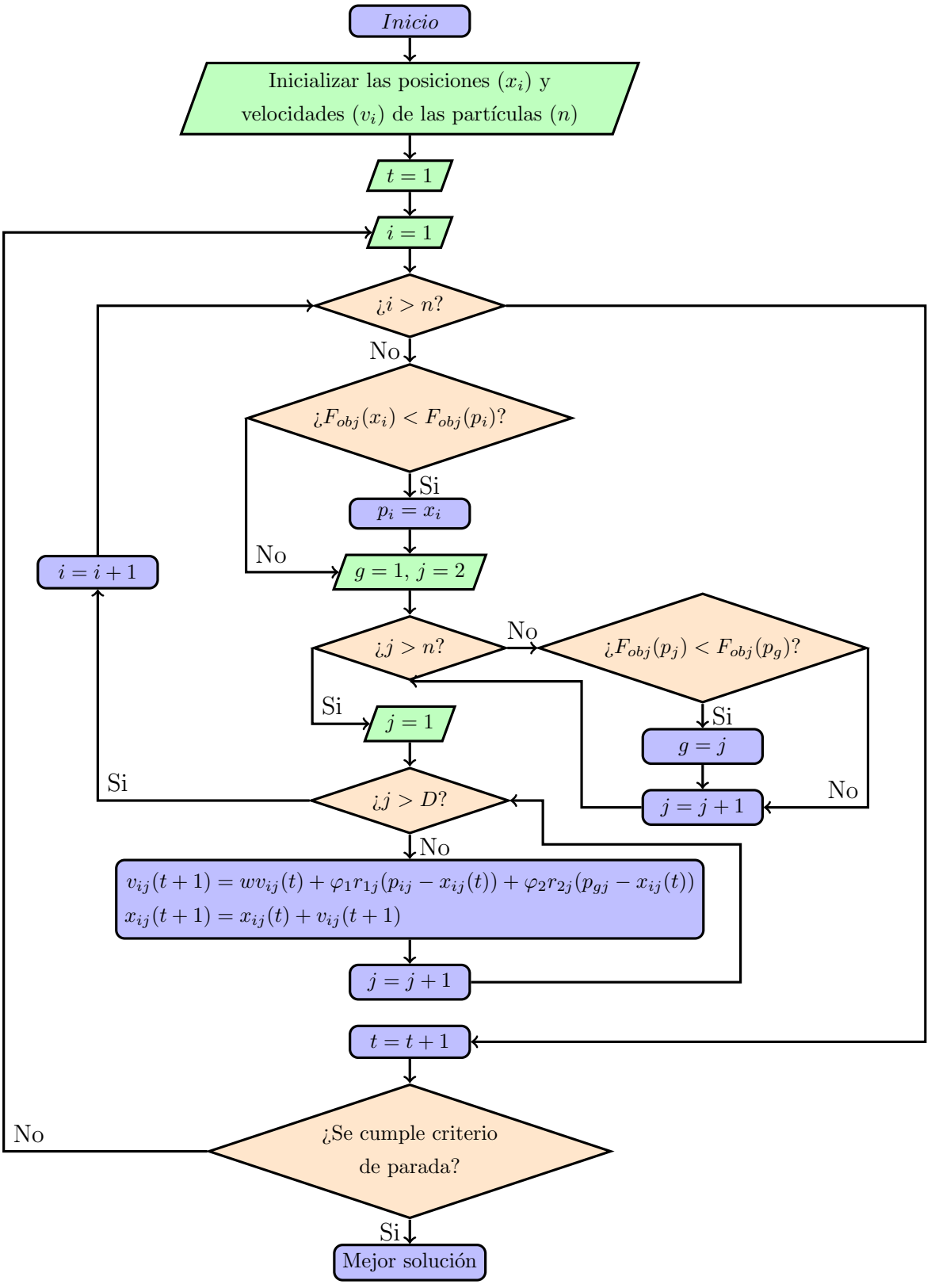


Figura B-4: Algoritmo del optimizador enjambre de partículas (PSO).

B.3.3. Algoritmo genético

Los métodos evolutivos corresponden a técnicas de búsqueda que mejoran su desempeño al muestrear regiones promisorias del espacio de búsqueda, mediante la aplicación de mecanismos de selección natural a la población de soluciones candidatas. Este tipo de algoritmos surgieron en la década de 1950 (Spall, 2005).

El algoritmo más conocido en esta categoría corresponde al algoritmo genético (GA), el cual tiene como idea central la generación aleatoria de una población inicial de soluciones, el cálculo del desempeño de dichas soluciones (evaluación de la función objetivo) y la aplicación de mecanismos evolutivos mediante operadores (recombinación, selección, mutación, inmigración) para producir la siguiente generación. El algoritmo genético posee algunas ventajas como versatilidad en la forma de la función objetivo, introducción directa de restricciones en las variables de optimización y posibilidad de paralelización (disminuye el tiempo de cómputo). El algoritmo básico de GA se muestra en la Figura **B-5**, en donde g es el contador de generaciones. La Tabla **B-3** presenta los parámetros que controlan el comportamiento del optimizador GA en el *global optmization toolbox* de Matlab® R2018.

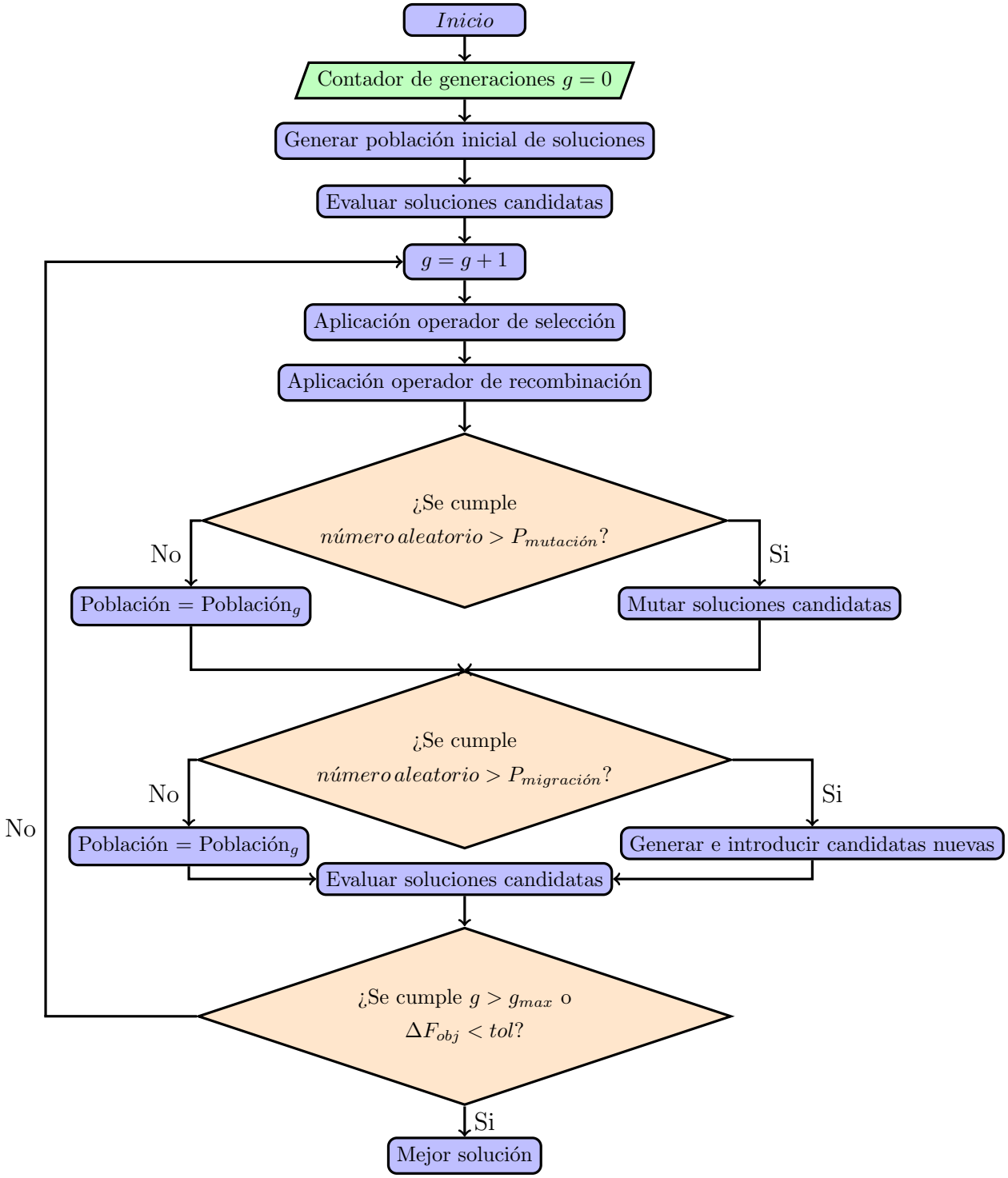


Figura B-5: Algoritmo del optimizador algoritmo genético (GA).

Tabla B-3: Descripción de los parámetros internos de GA.

`PopulationSize`	Especifica el número de individuos que existen en cada generación.		
`SelectionFcn`	Especifica la selección de parentales para la próxima generación	`selectionstochunif`	Genera una linea entre los valores de los padres con secciones proporcionales al valor escalado de cada uno.
		`selectionremainder`	Clasifica los parentales por el valor de su entero y aplica selección por ruleta, con probabilidad equivalente a su valor escalado.
		`selectionuniform`	Selecciona los parentales al azar con una distribución uniforme.
		`selectionroulette`	Selecciona parentales por el método de ruleta, con probabilidad equivalente a su desempeño.
		`selectiontournament`	Selecciona los parental por selección aleatoria de un numero de "jugadores" y elección de aquel con mejor desempeño.
`MutationFcn`	Controla la aparición de mutaciones en los individuos	`mutationgaussian`	agrega un número aleatorio de una distribución Gaussiana con media 0. La desviación estándar esta determinada por los parámetros `scale` y `shrink`.
		`mutationuniform`	Primero realiza selección de una fracción de los atributos del individuo, luego cada atributo tiene una probabilidad `Mrate` de ser mutado.
`CrossoverFraction`	Fracción de individuos que son generados por entrecruzamiento en la próxima generación.		
`CrossoverFcn`	Controla el entrecruzamiento entre parentales	`crossoverscattered`	Intercambia aleatoriamente atributos de los parentales
		`crossoersinglepoint`	Selecciona un número aleatorio de atributos, posteriormente, los combina este numero de atributos del parental 1 con el resto del parental 2
		`crossoverintermediate`	Genera un individuo por ponderación de los atributos de los parentales, con peso igual a `CIratio` más cercano a aquel con mejor desempeño.
		`crossoverheuristic`	Genera un individuo por ponderación de los atributos de los parentales, con peso igual a `CHratio` más cercano a aquel con peor desempeño.
		`crossoverarithmetic`	Genera un individuo cuyos atributos corresponden a la media aritmética de los atributos de los parentales.
`MigrationDirection`	Controla el movimiento de individuos entre subpoblaciones	`forward`	La migración se da hacia la siguiente subpoblación.
		`both`	La subpoblación actual migra hacia las subpoblaciones anterior y siguiente
`MigrationFraction`	Especifica la fracción de la subpoblación actual que va a migrar.		
`MigrationInterval`	Especifica cuantas generaciones se dan antes de que ocurra la migración.		

B.4. Sintonización de algoritmos de optimización

B.4.1. Templado simulado

Normalización por valor mínimo

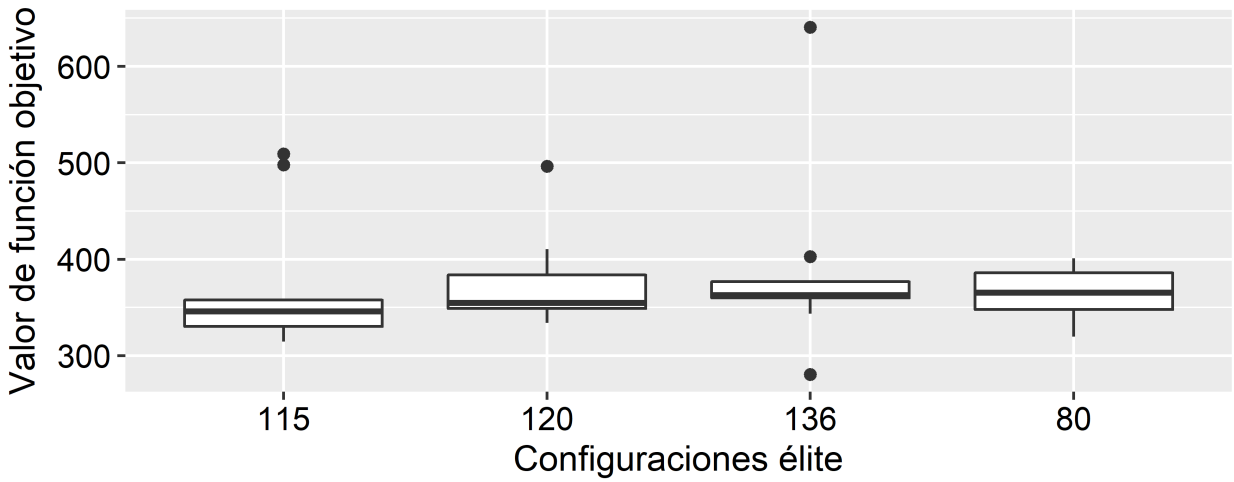


Figura B-6: Dispersión de valores de función objetivo para las configuraciones élite del optimizador SA con normalización por valor mínimo.

Normalización por valor máximo

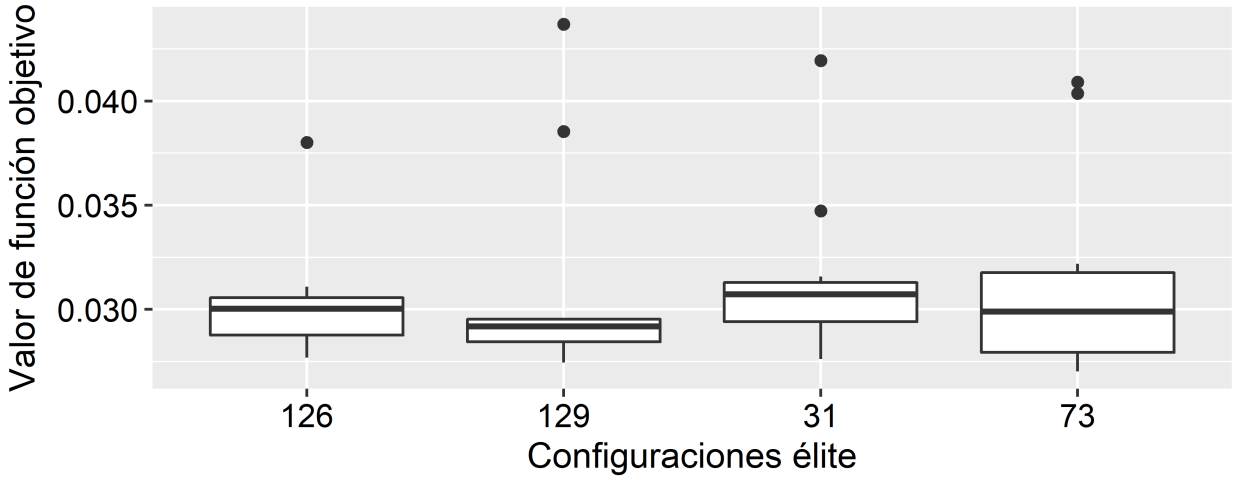


Figura B-7: Dispersión de valores de función objetivo para las configuraciones élite del optimizador SA con normalización por valor máximo.

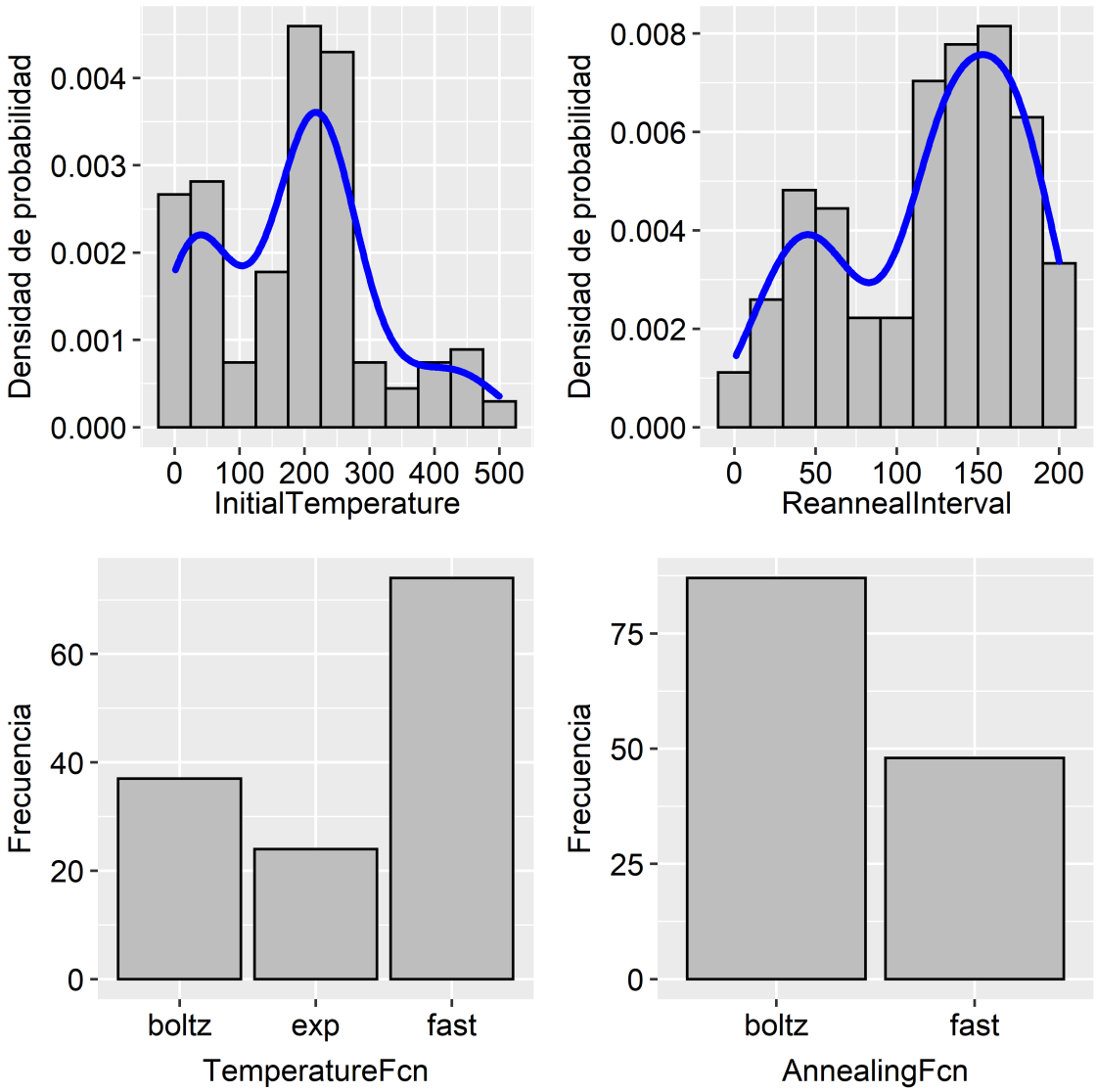


Figura B-8: Diagrama de frecuencia para los parámetros de SA con normalización por valor máximo para `InitialTemperature`, `RannealInterval`, `TemperatureFcn` y `AnnealingFcn`.

Normalización por valor medio

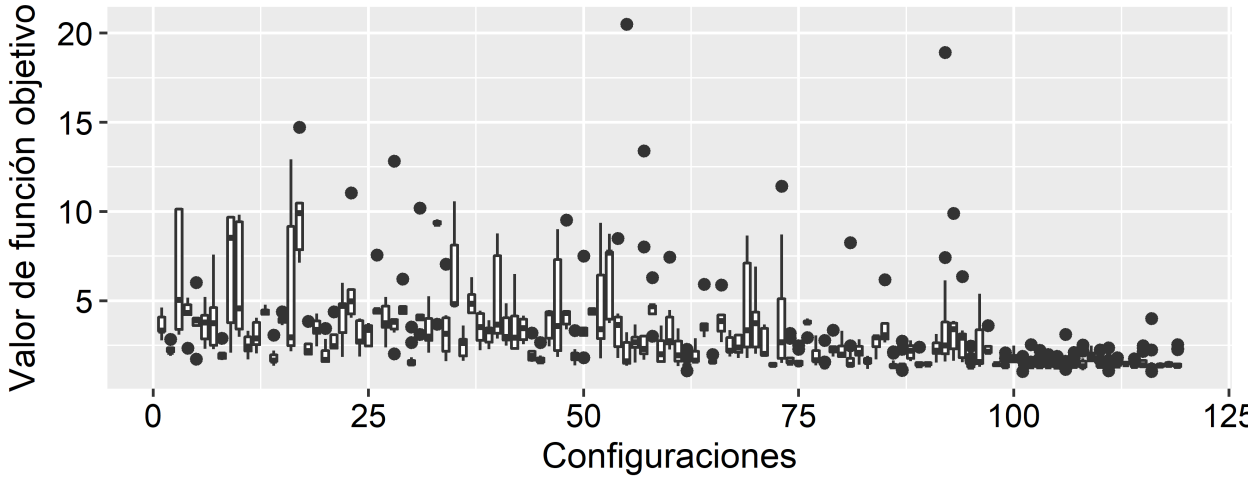


Figura B-9: Dispersión de valores de función objetivo para todas las configuraciones generadas del optimizador SA con normalización por valor medio.

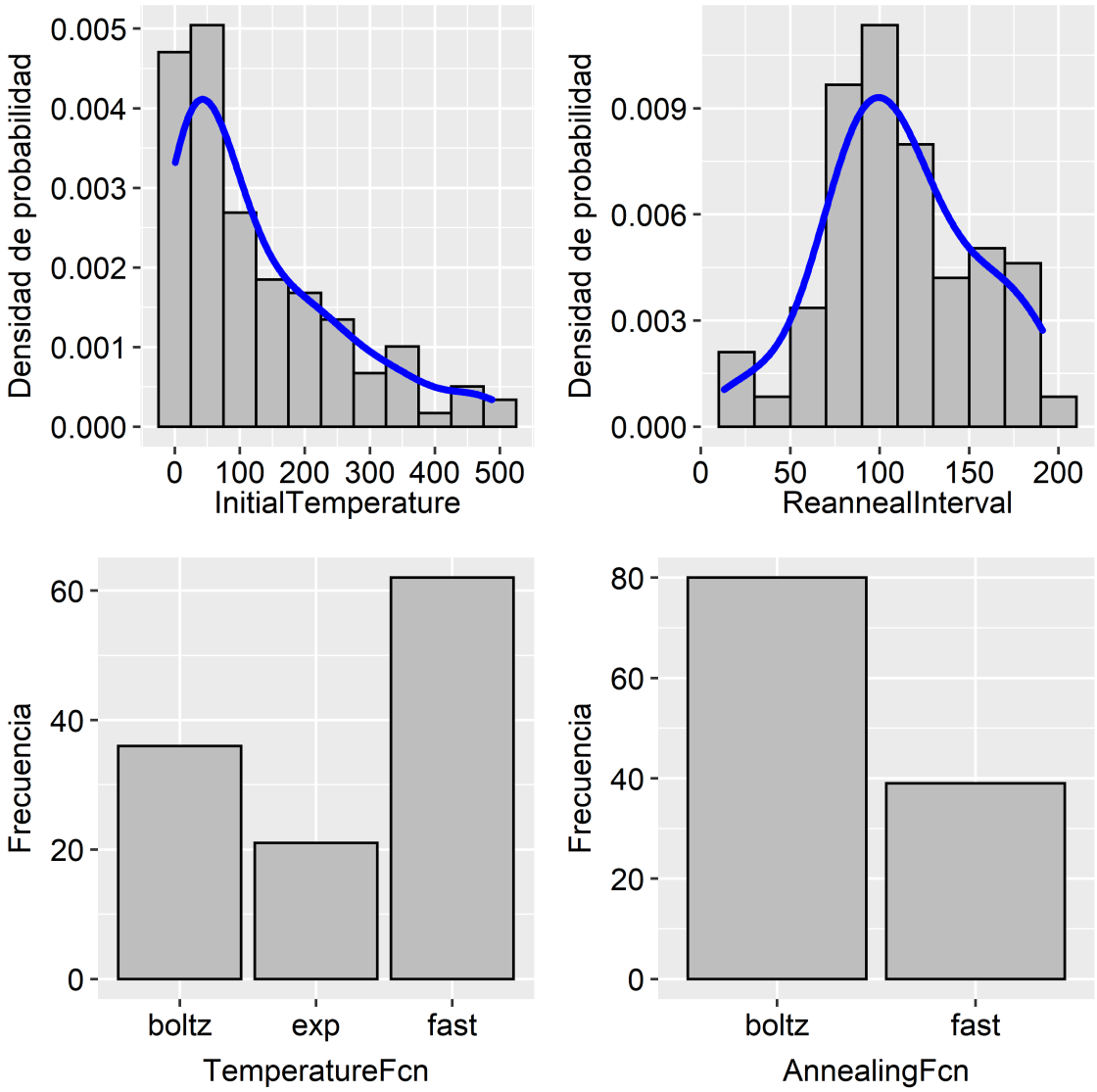


Figura B-10: Diagrama de frecuencia para los parámetros de SA con normalización por valor medio para `InitialTemperature`, `RannealInterval`, `TemperatureFcn` y `AnnealingFcn`.

B.4.2. Enjambre de partículas

Normalización por valor máximo

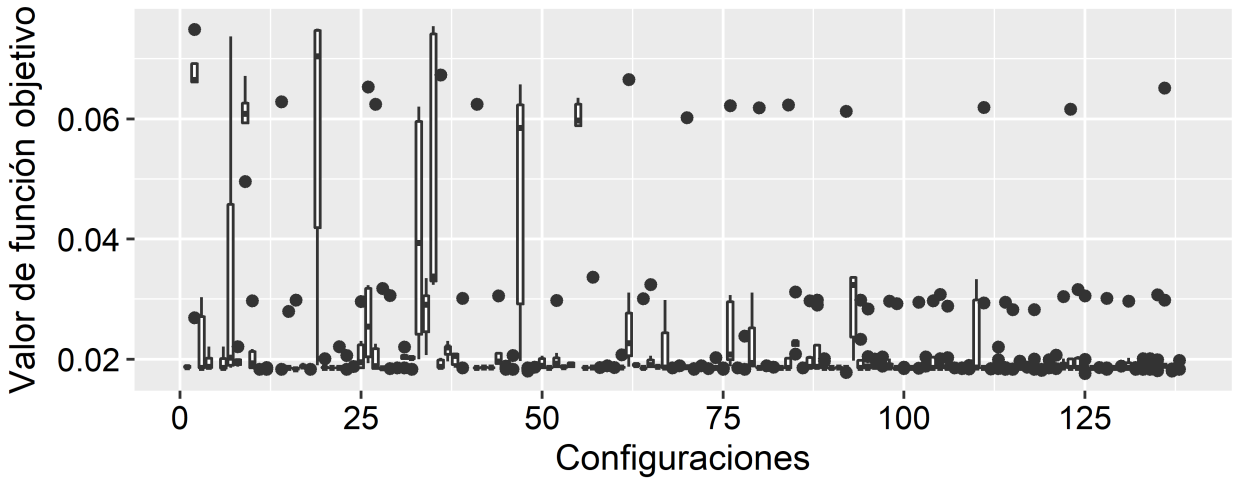


Figura B-11: Dispersión de valores de función objetivo para todas las configuraciones generadas del optimizador PSO con normalización por valor máximo.

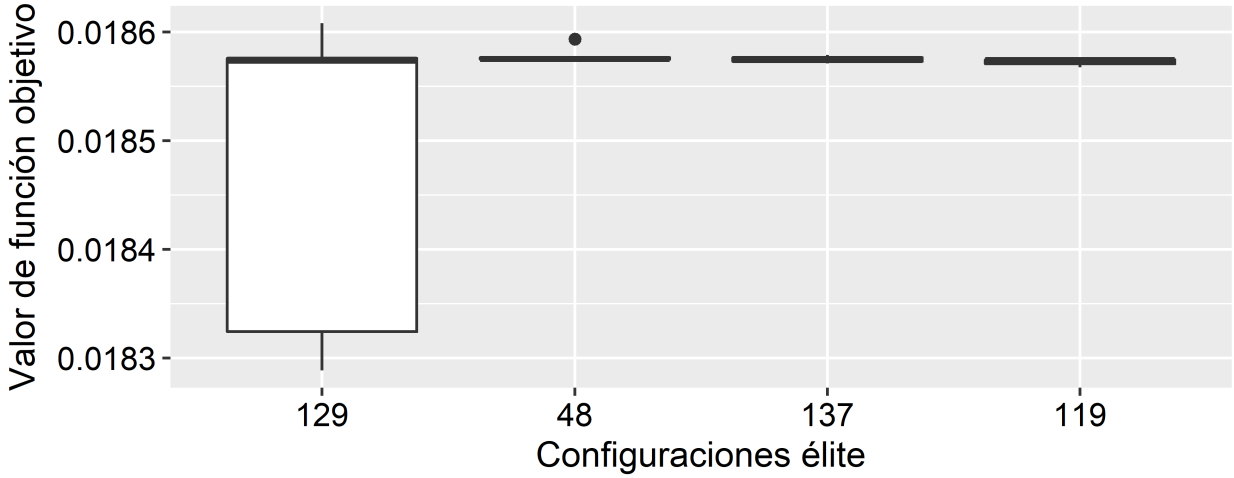


Figura B-12: Dispersión de valores de función objetivo para las configuraciones élite del optimizador PSO con normalización por valor máximo.

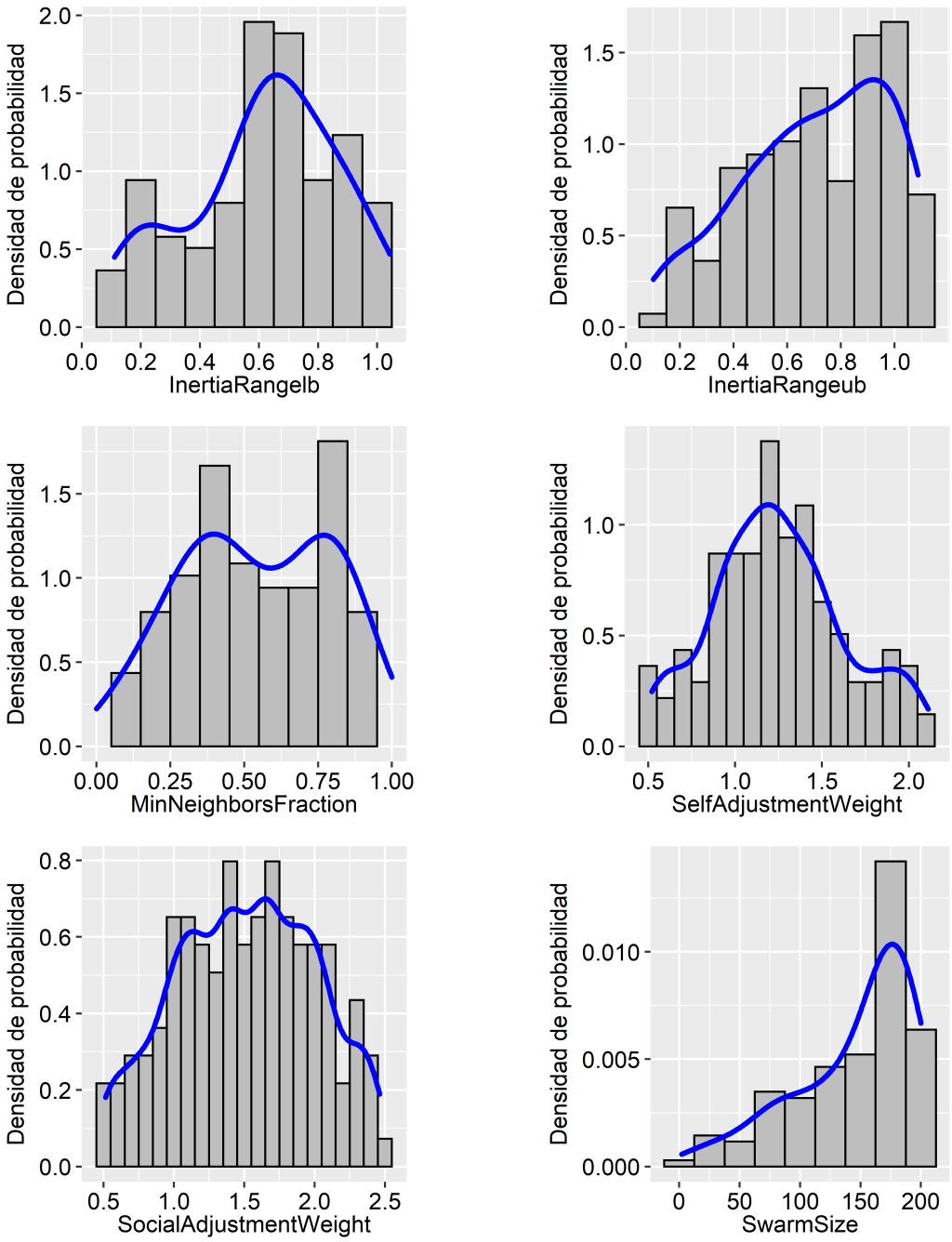


Figura B-13: Diagrama de frecuencia para los parámetros de PSO con normalización por valor máximo para `InertiaRangelb`, `InertiaRangeub`, `MinNeighborsFraction`, `SelfAdjustmentWeight`, `SocialAdjusmentWeight` y `SwarmSize`.

Normalización por valor medio

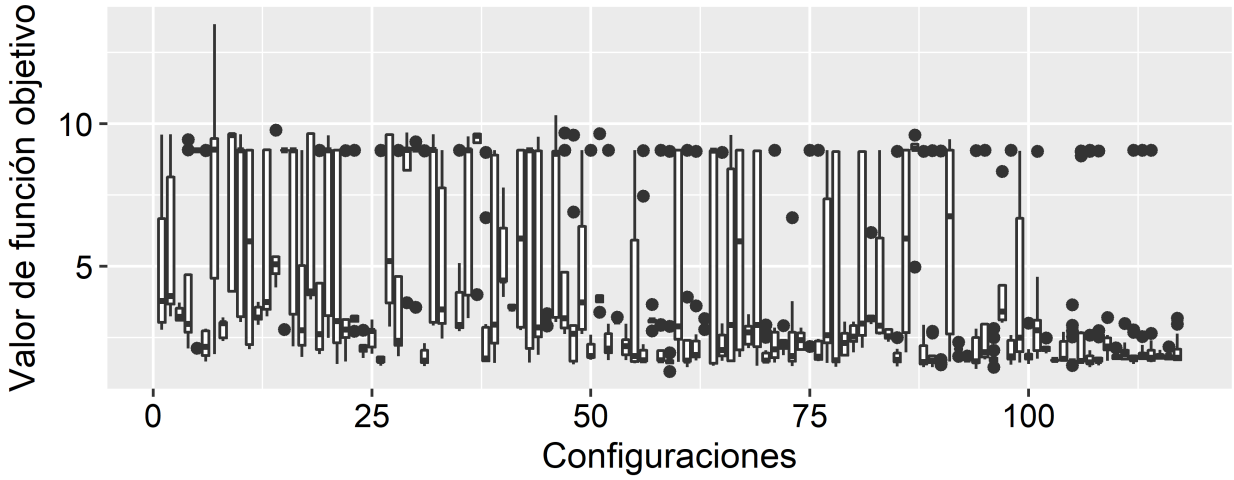


Figura B-14: Dispersión de valores de función objetivo para todas las configuraciones generadas del optimizador PSO con normalización por valor medio.

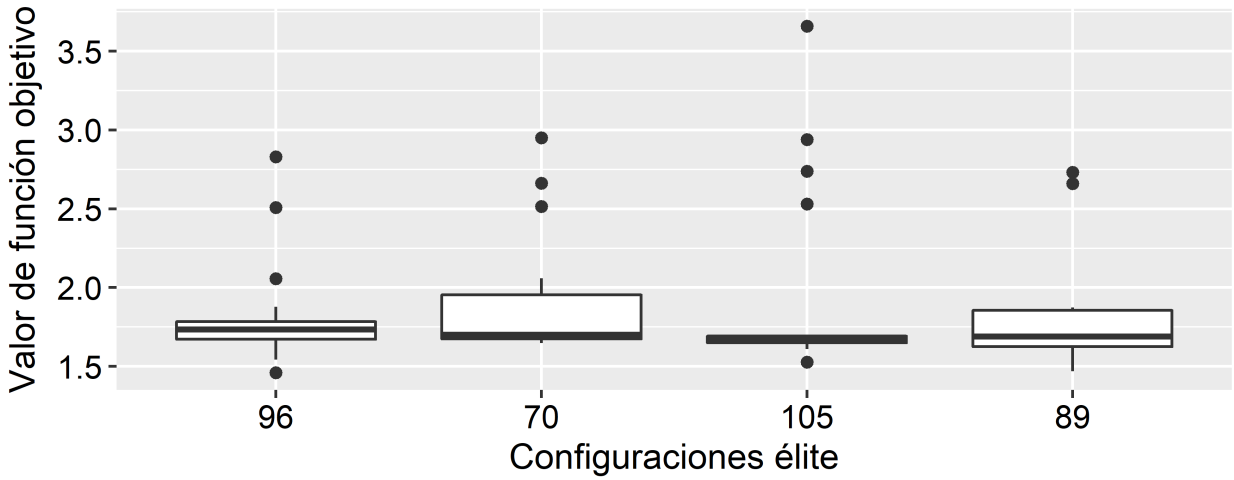


Figura B-15: Dispersión de valores de función objetivo para las configuraciones élite del optimizador PSO con normalización por valor medio.

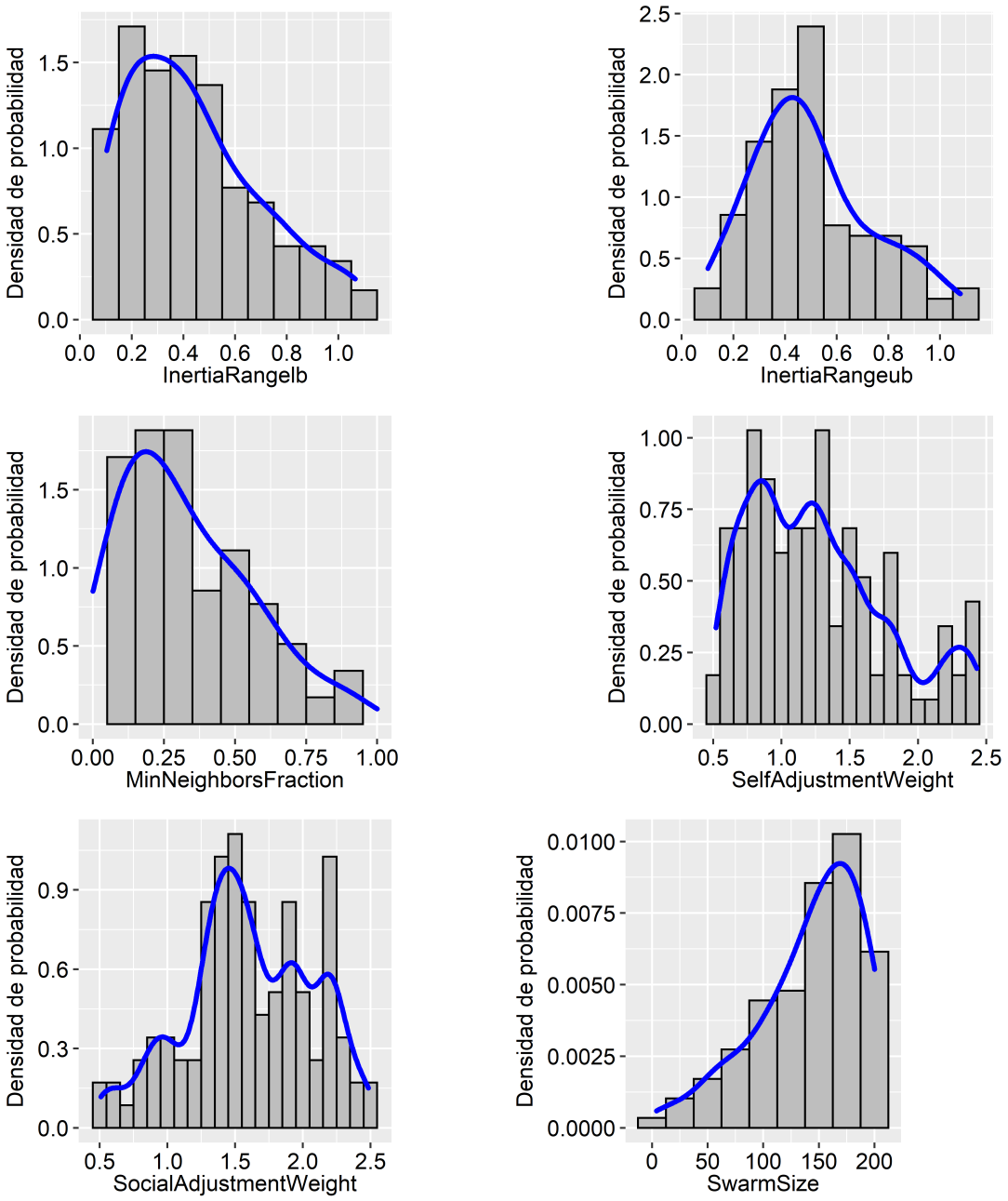


Figura B-16: Diagrama de frecuencia para los parámetros de PSO con normalización por valor medio para `InertiaRangelb`, `InertiaRangeub`, `MinNeighborsFraction`, `SelfAdjustmentWeight`, `SocialAdjusmentWeight` y `SwarmSize`.

Normalización por valor mínimo

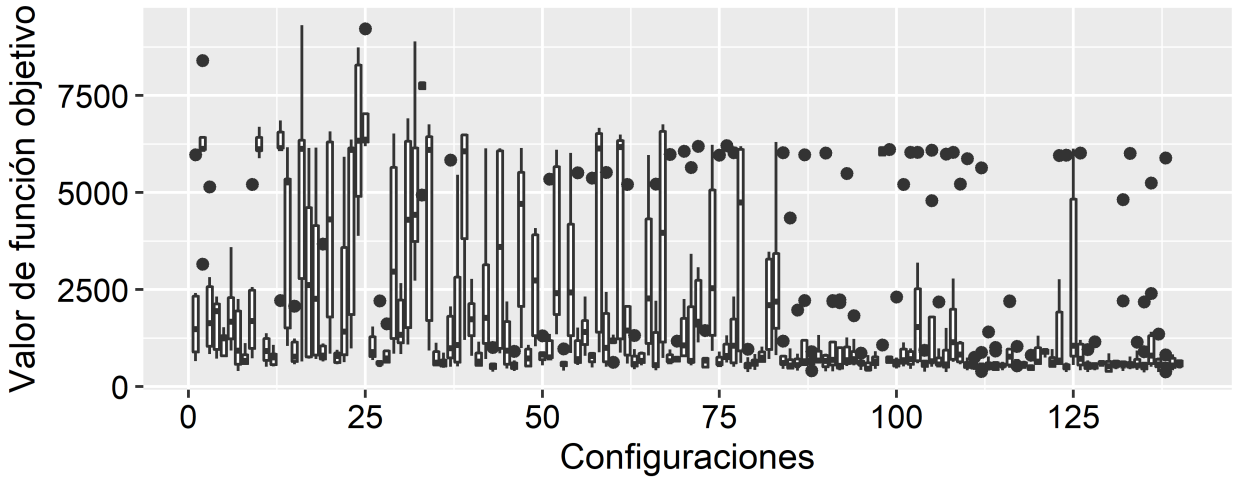


Figura B-17: Dispersión de valores de función objetivo para todas las configuraciones generadas del optimizador PSO con normalización por valor mínimo.

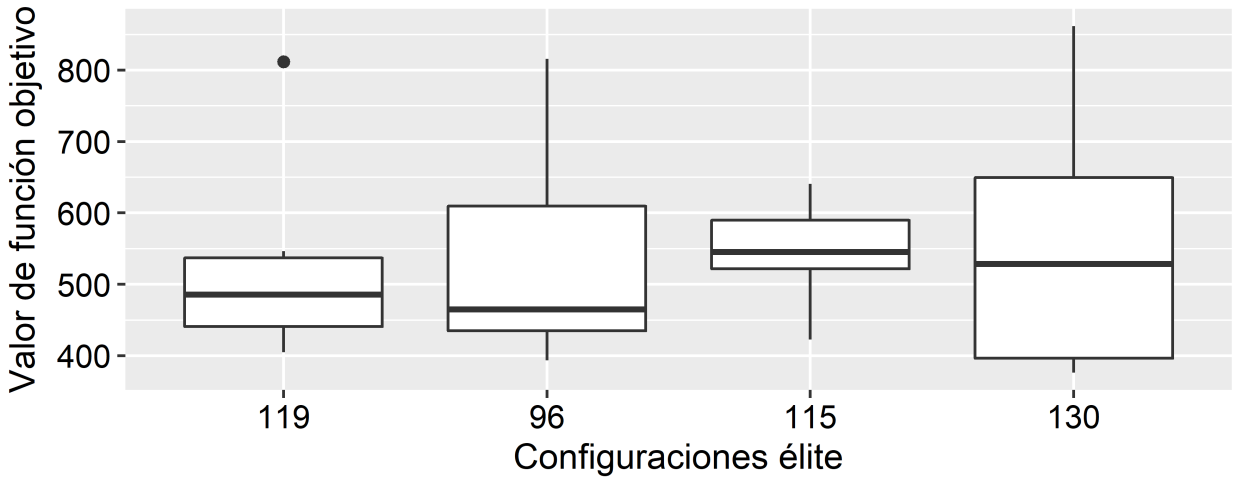


Figura B-18: Dispersión de valores de función objetivo para las configuraciones élite del optimizador PSO con normalización por valor medio.

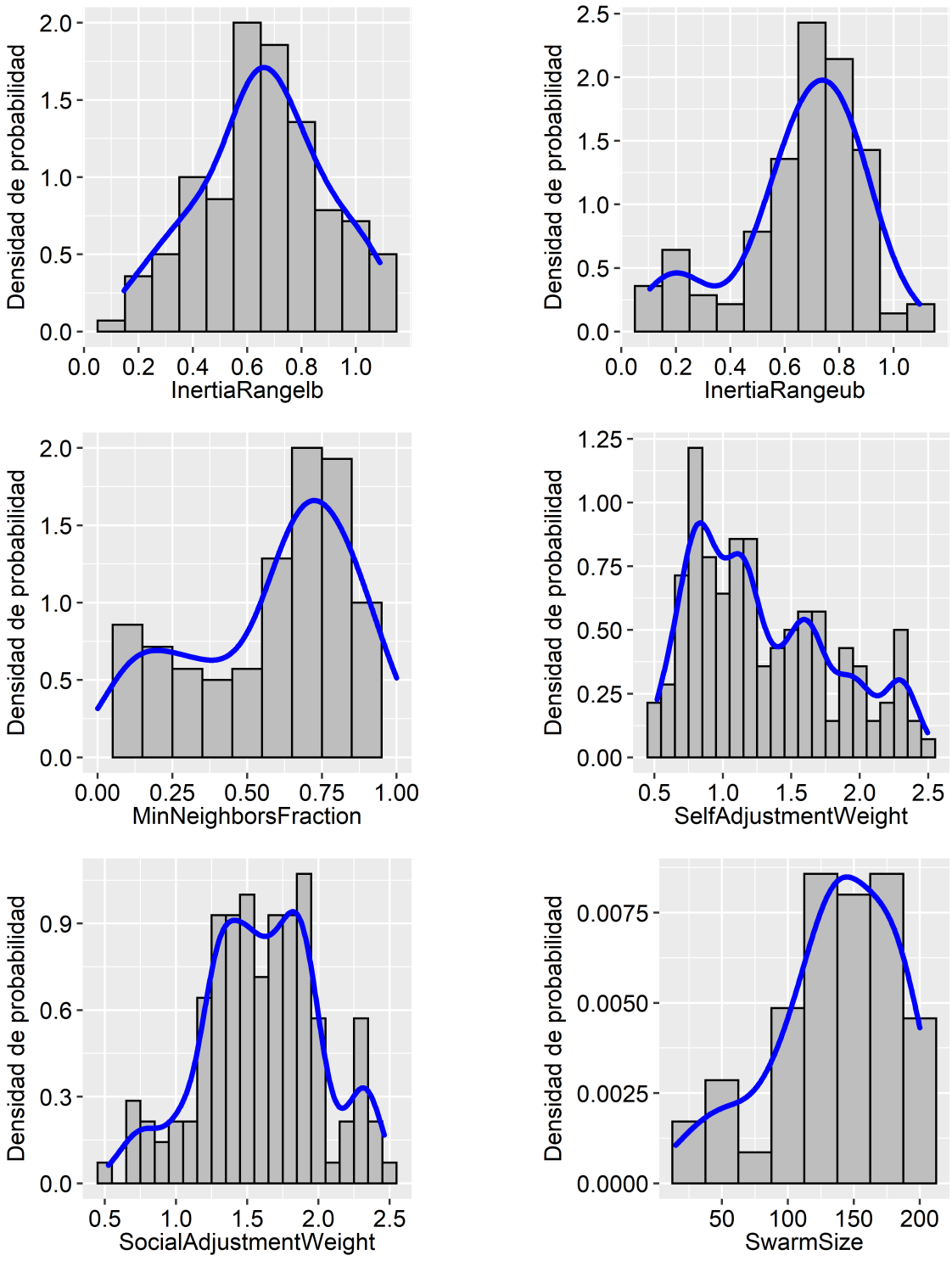


Figura B-19: Diagrama de frecuencia para los parámetros de PSO con normalización por valor mínimo para `InertiaRangelb`, `InertiaRangeub`, `MinNeighborsFraction`, `SelfAdjustmentWeight`, `SocialAdjusmentWeight` y `SwarmSize`.

B.4.3. Algoritmo genético

Normalización por valor máximo

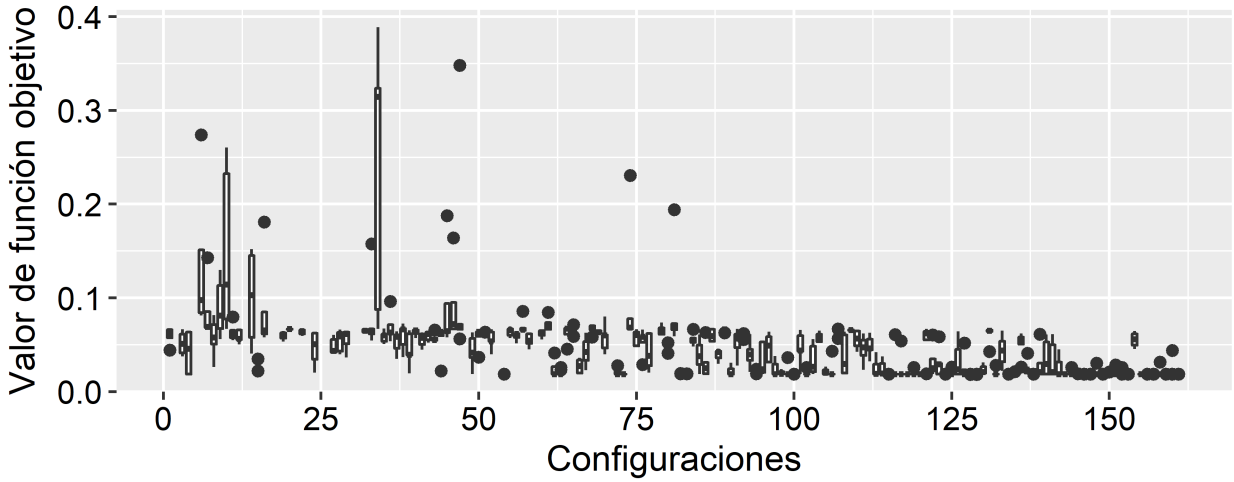


Figura B-20: Dispersión de valores de función objetivo para todas las configuraciones generadas del optimizador GA con normalización por valor máximo.

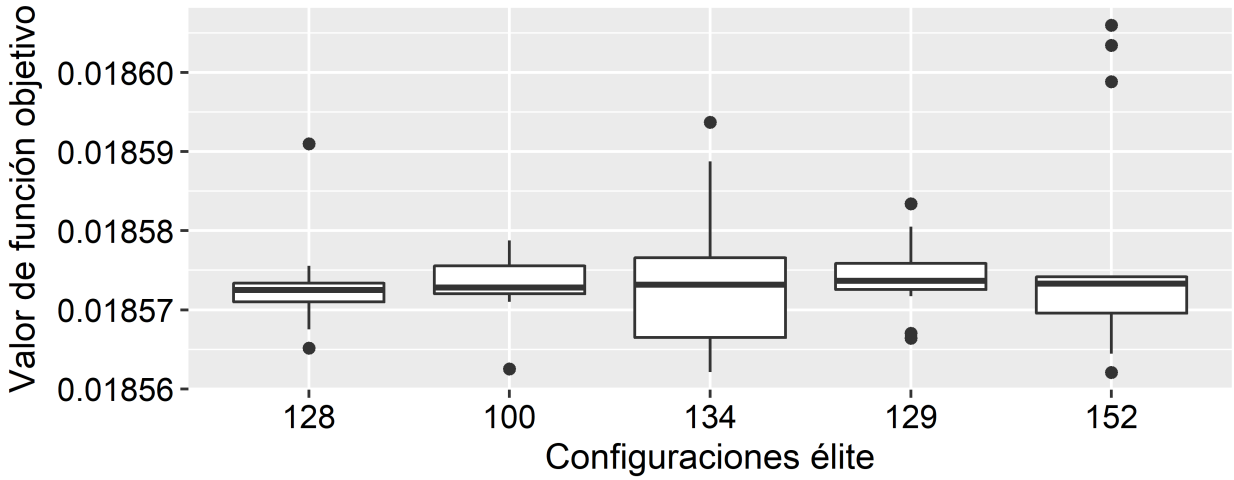


Figura B-21: Dispersión de valores de función objetivo para las configuraciones élite del optimizador GA con normalización por valor mínimo.

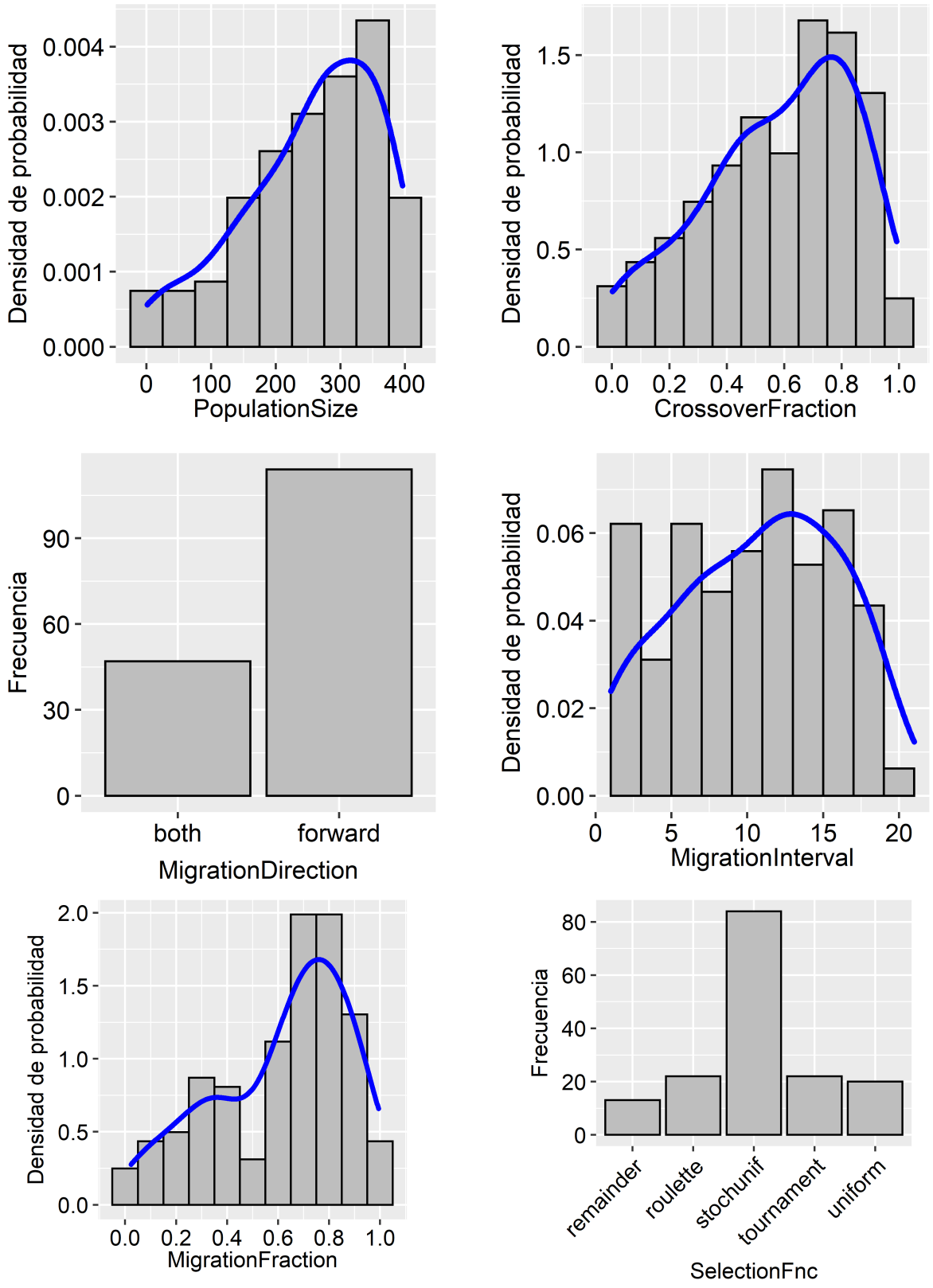


Figura B-22: Diagrama de frecuencia para los parámetros de GA con normalización por valor máximo para `PopulationSize`, `CrossoverFraction`, `MigrationDirection`, `MigrationInterval`, `MigrationFraction` y `SelectionFnc`.

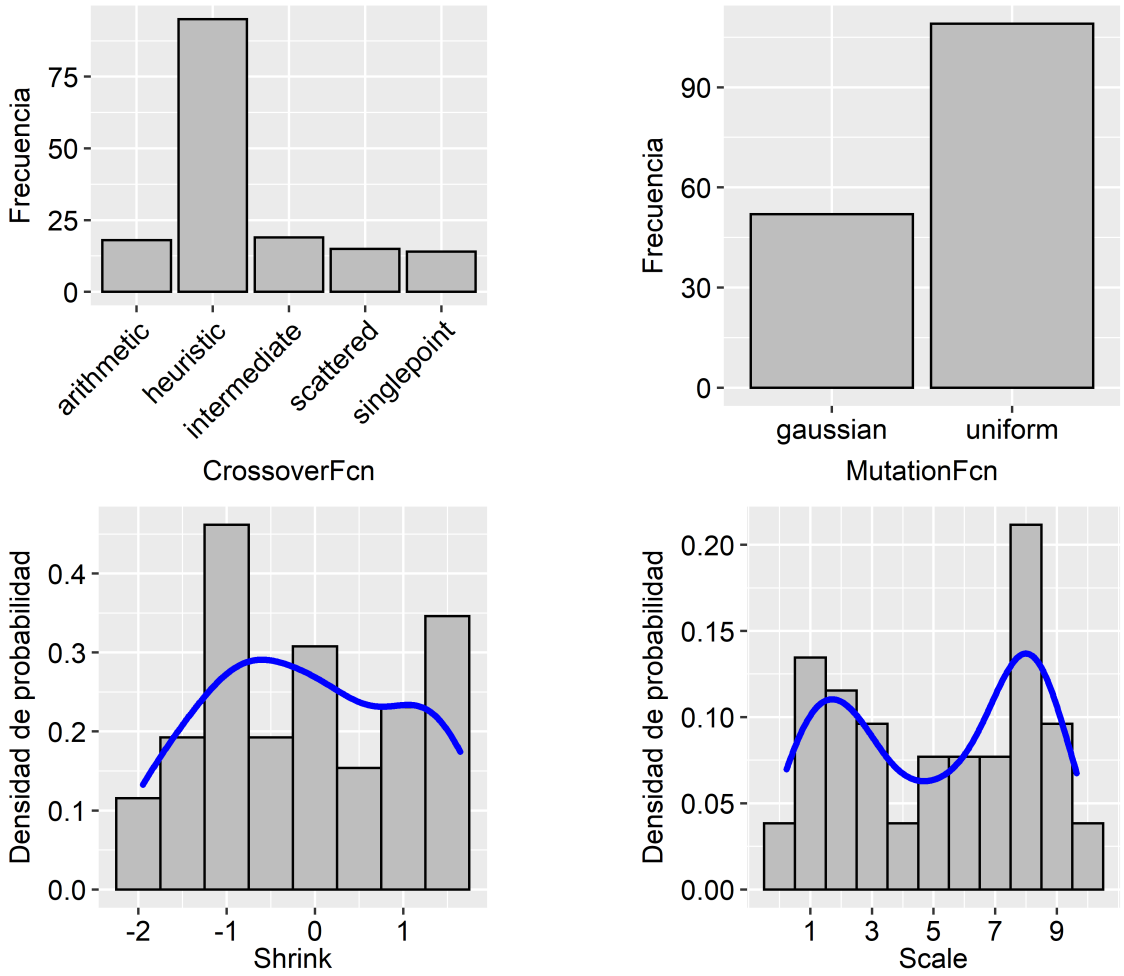


Figura B-23: Diagrama de frecuencia para los parámetros de GA con normalización por valor máximo para `CrossoverFnc`, `MutationFnc`, `Shrink` y `Scale`.

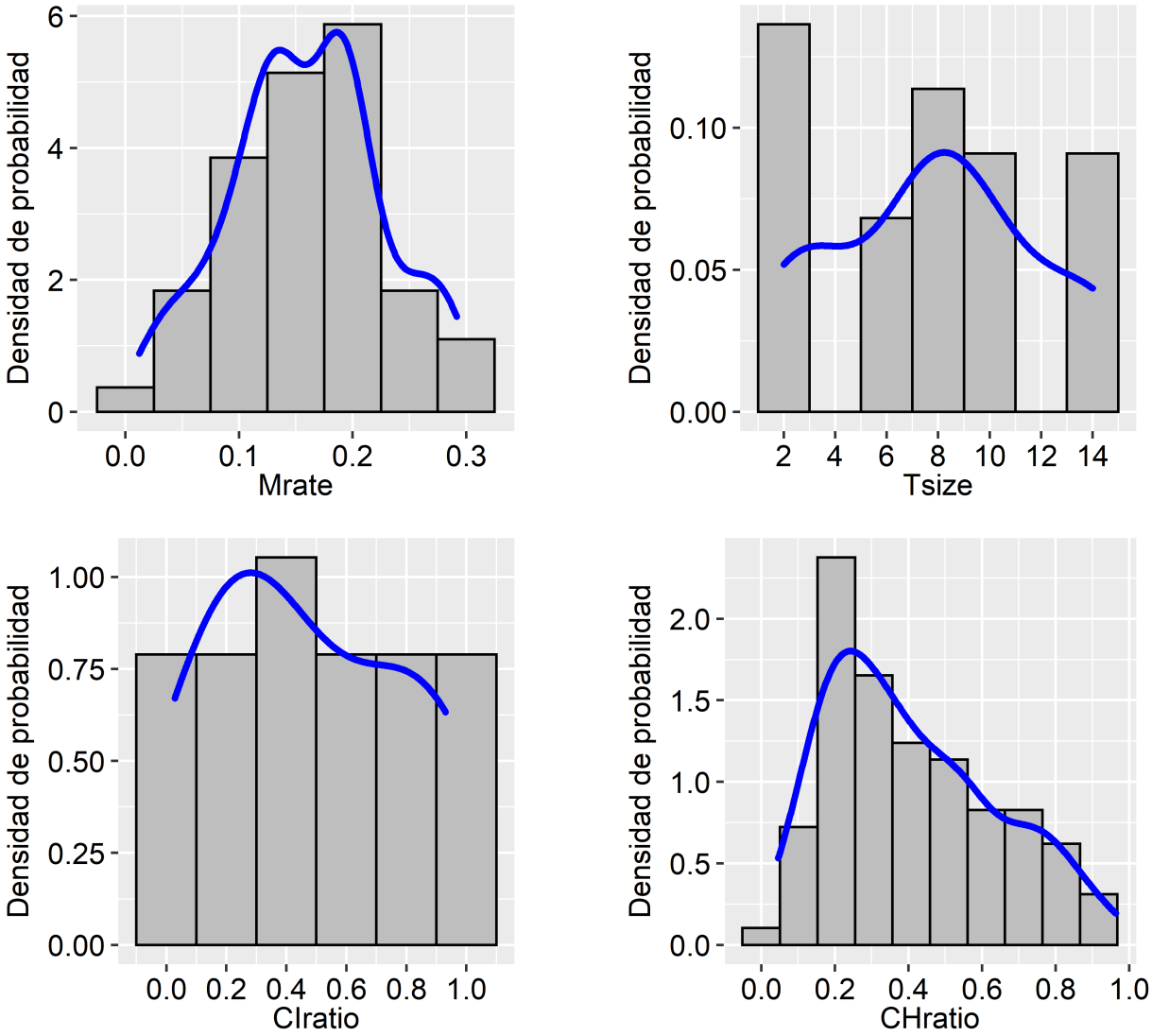


Figura B-24: Diagrama de frecuencia para los parámetros de GA con normalización por valor máximo para `Mrate`, `Tsize`, `CIratio` y `CHratio`.

Normalización por valor medio

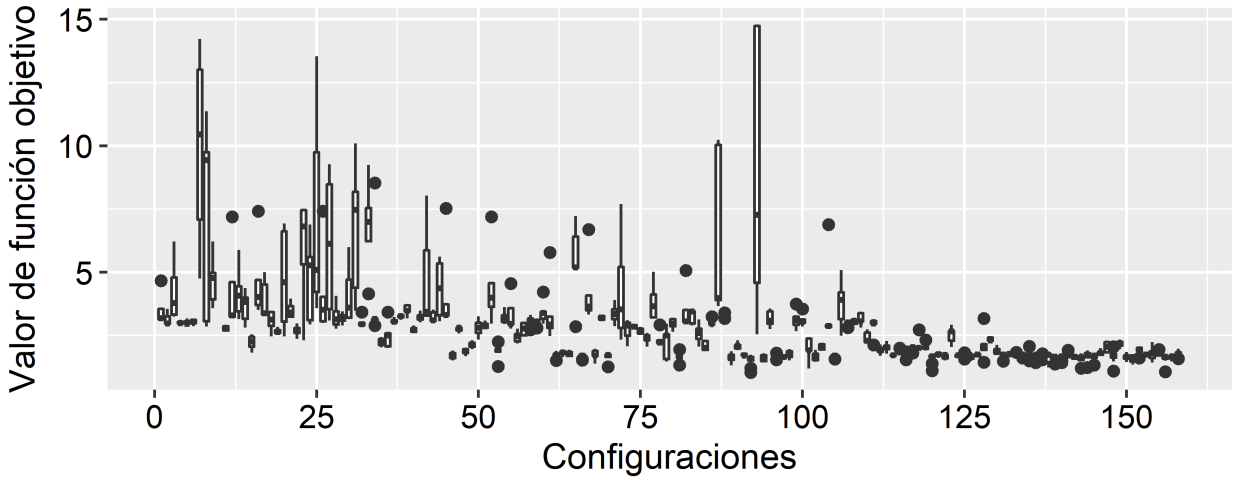


Figura B-25: Dispersión de valores de función objetivo para todas las configuraciones generadas del optimizador GA con normalización por valor medio.

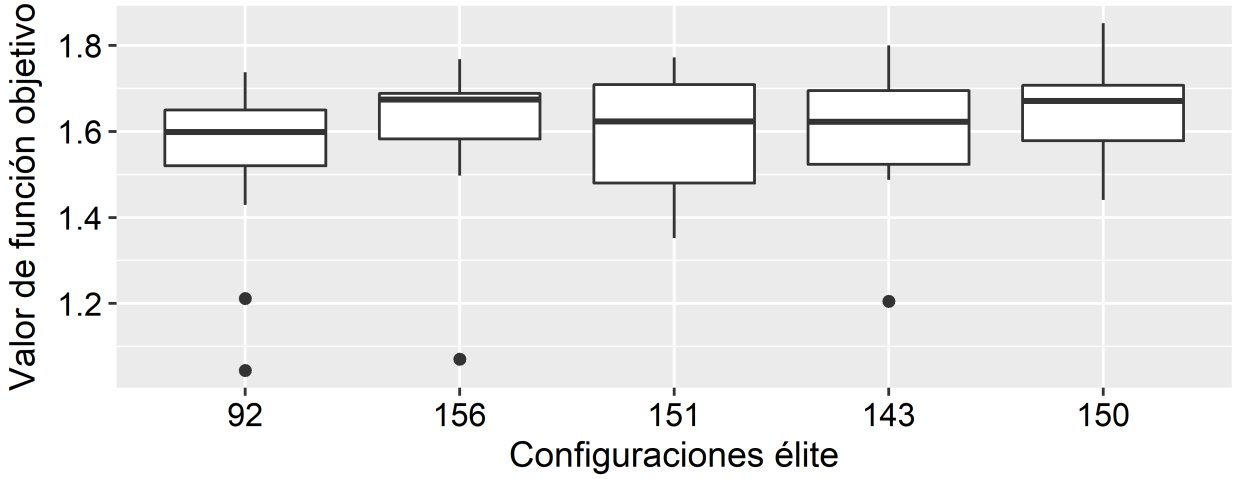


Figura B-26: Dispersión de valores de función objetivo para las configuraciones élite del optimizador GA con normalización por valor medio.

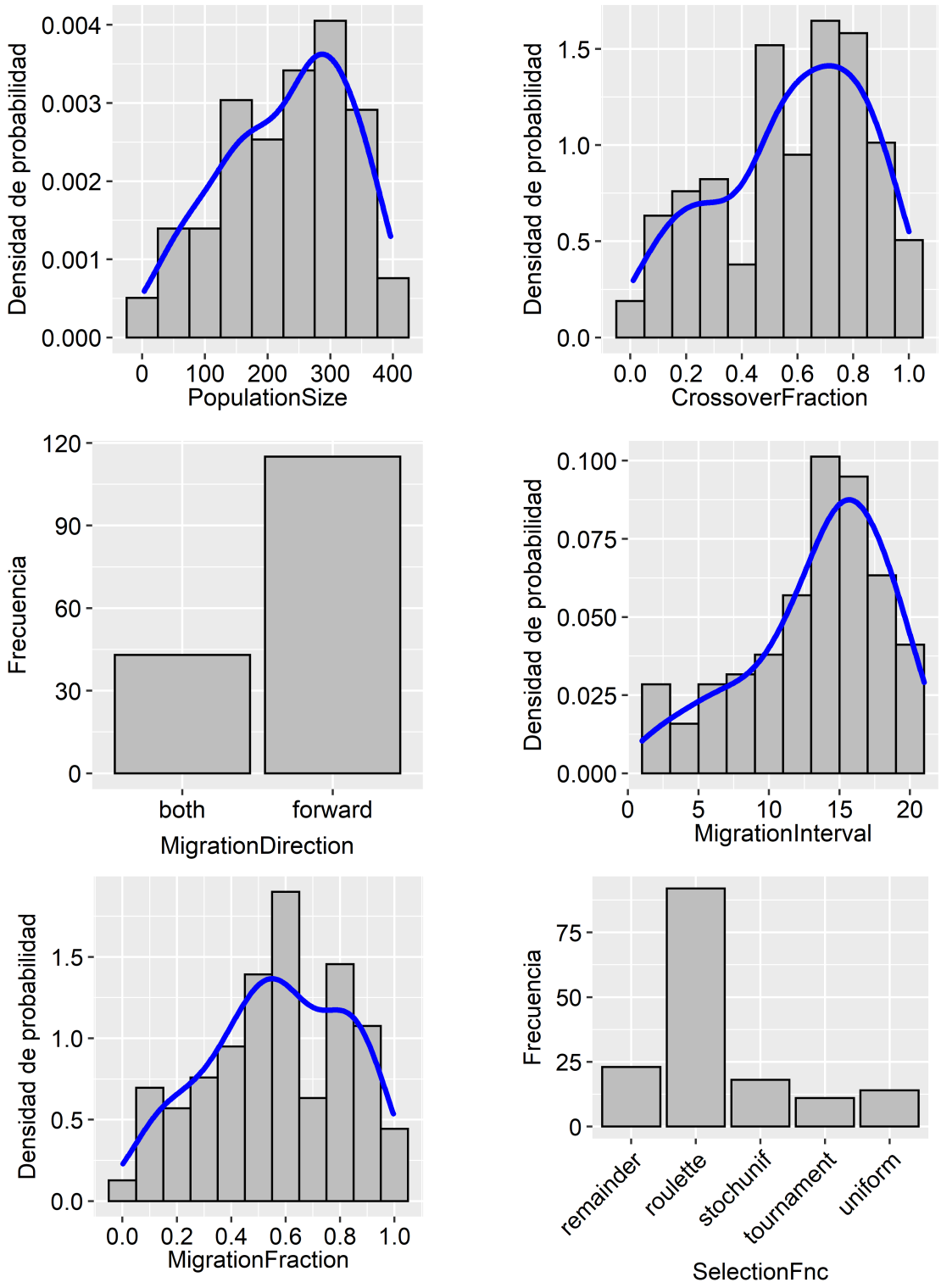


Figura B-27: Diagrama de frecuencia para los parámetros de GA con normalización por valor medio para `PopulationSize`, `CrossoverFraction`, `MigrationDirection`, `MigrationInterval`, `CrossoverFnc` y `MutationFnc`.

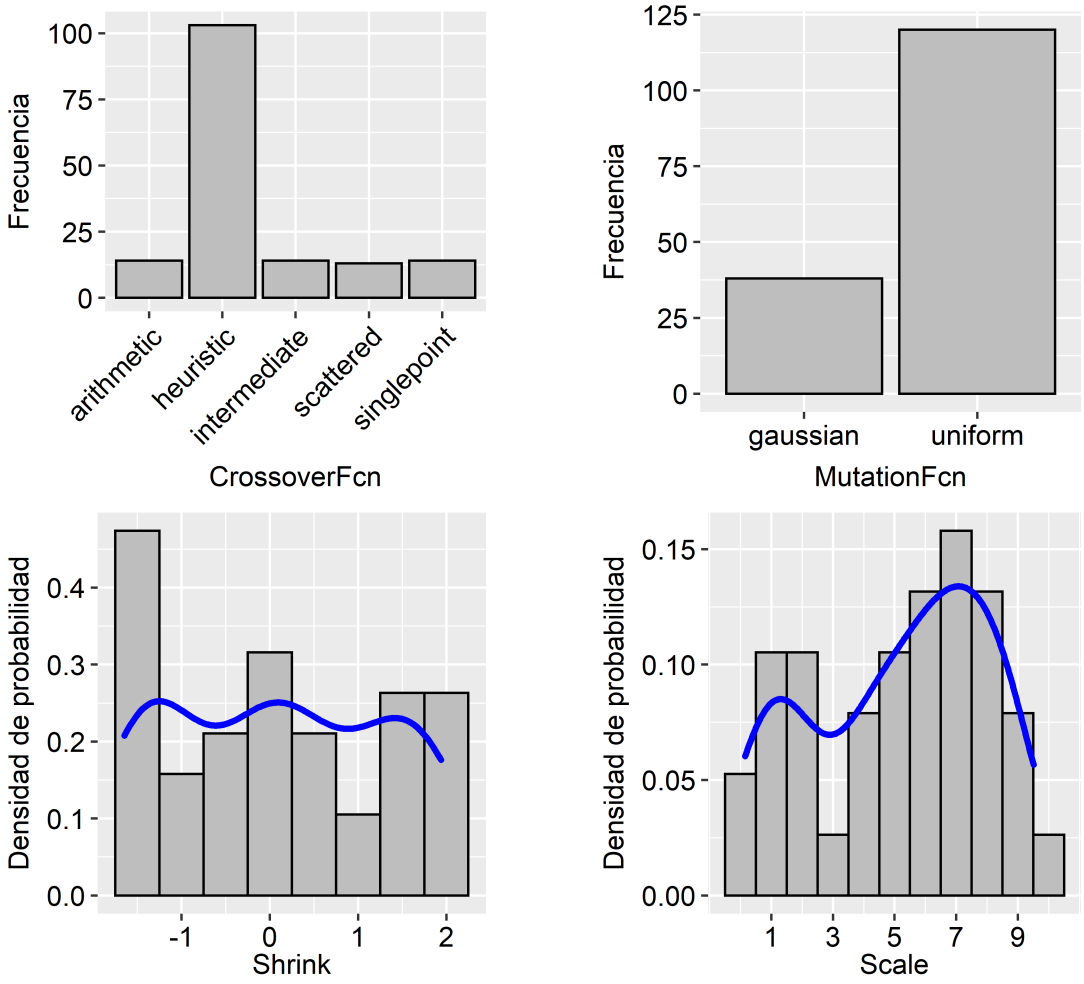


Figura B-28: Diagrama de frecuencia para los parámetros de GA con normalización por valor medio para `Shrink`, `Scale`, `Mrate` y `Tsize`.

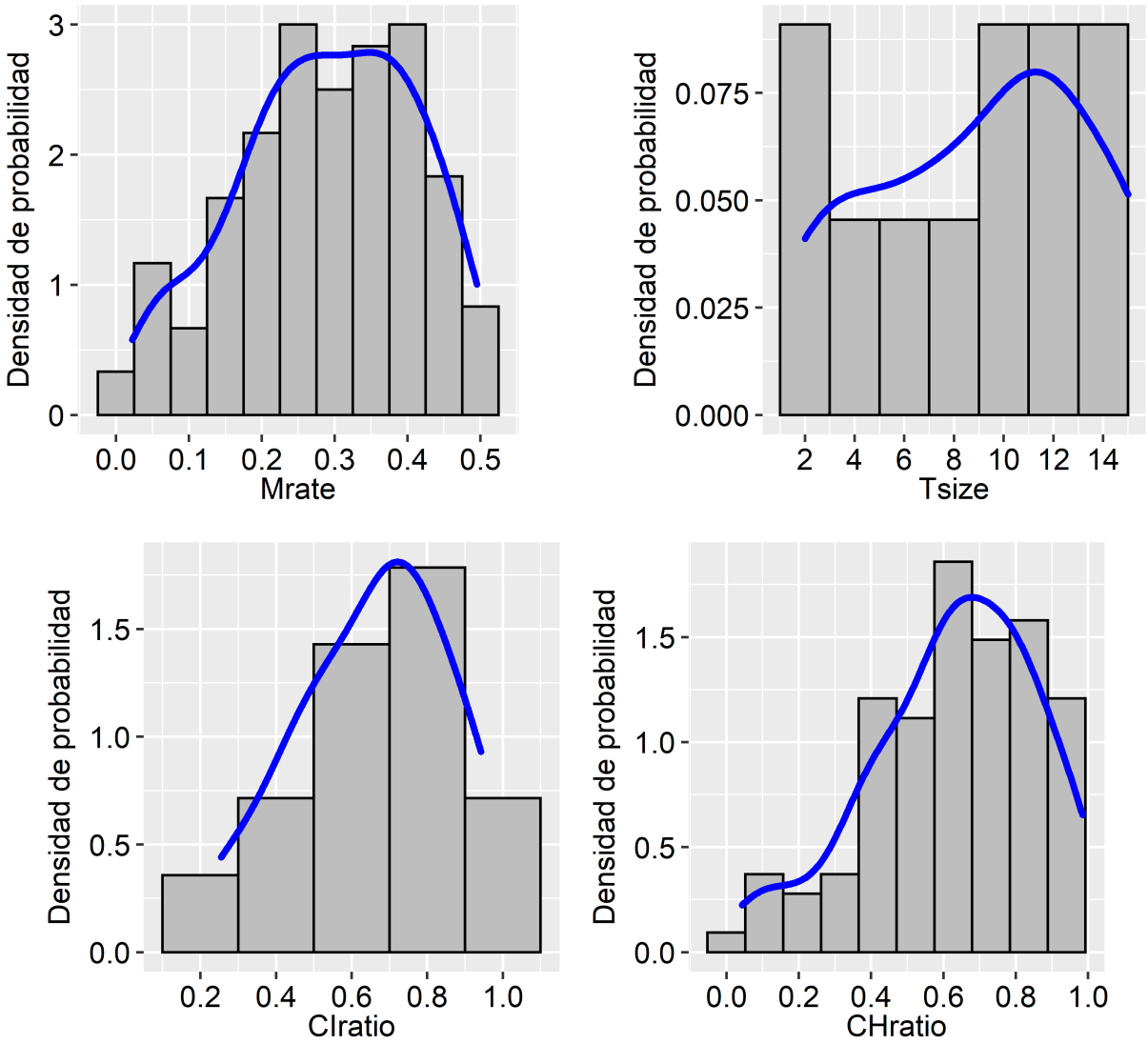


Figura B-29: Diagrama de frecuencia para los parámetros de GA con normalización por valor medio para `MigrationFraction`, `SelectionFnc`, `CIratio` y `CHratio`.

Normalización por valor mínimo

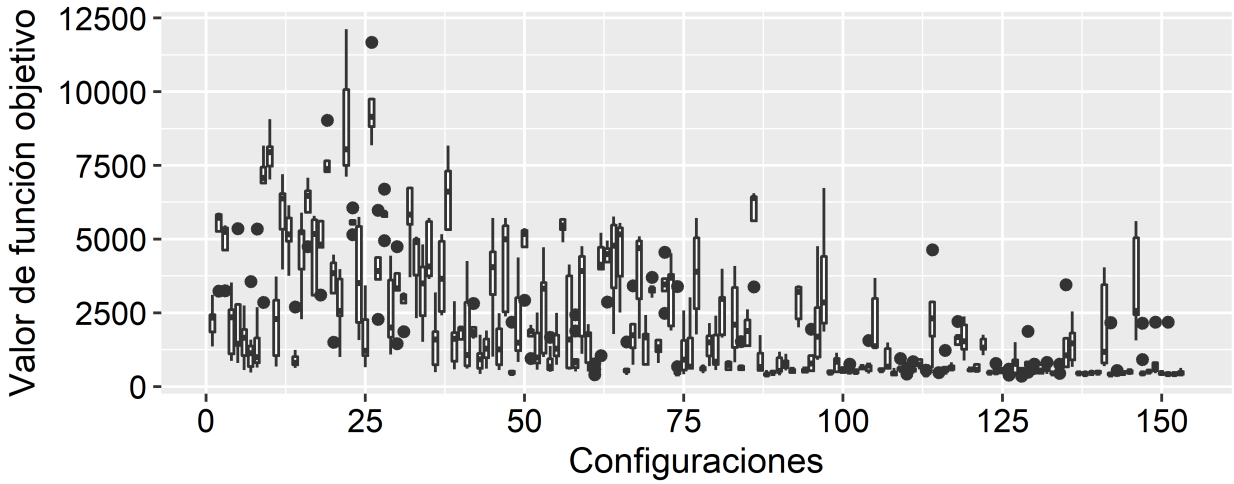


Figura B-30: Dispersión de valores de función objetivo para todas las configuraciones generadas del optimizador GA con normalización por valor mínimo.

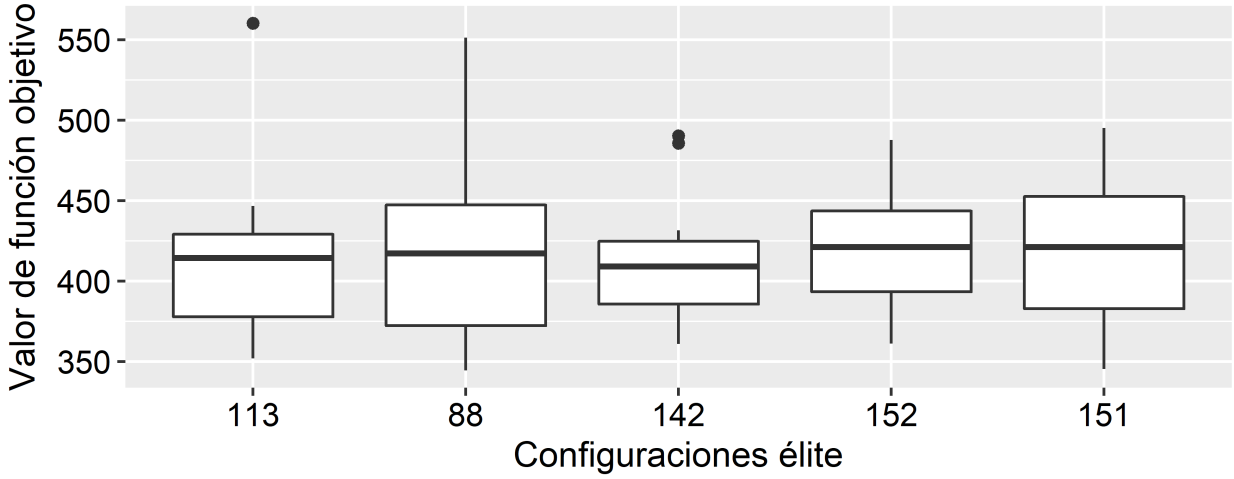


Figura B-31: Dispersión de valores de función objetivo para las configuraciones élite del optimizador GA con normalización por valor mínimo.

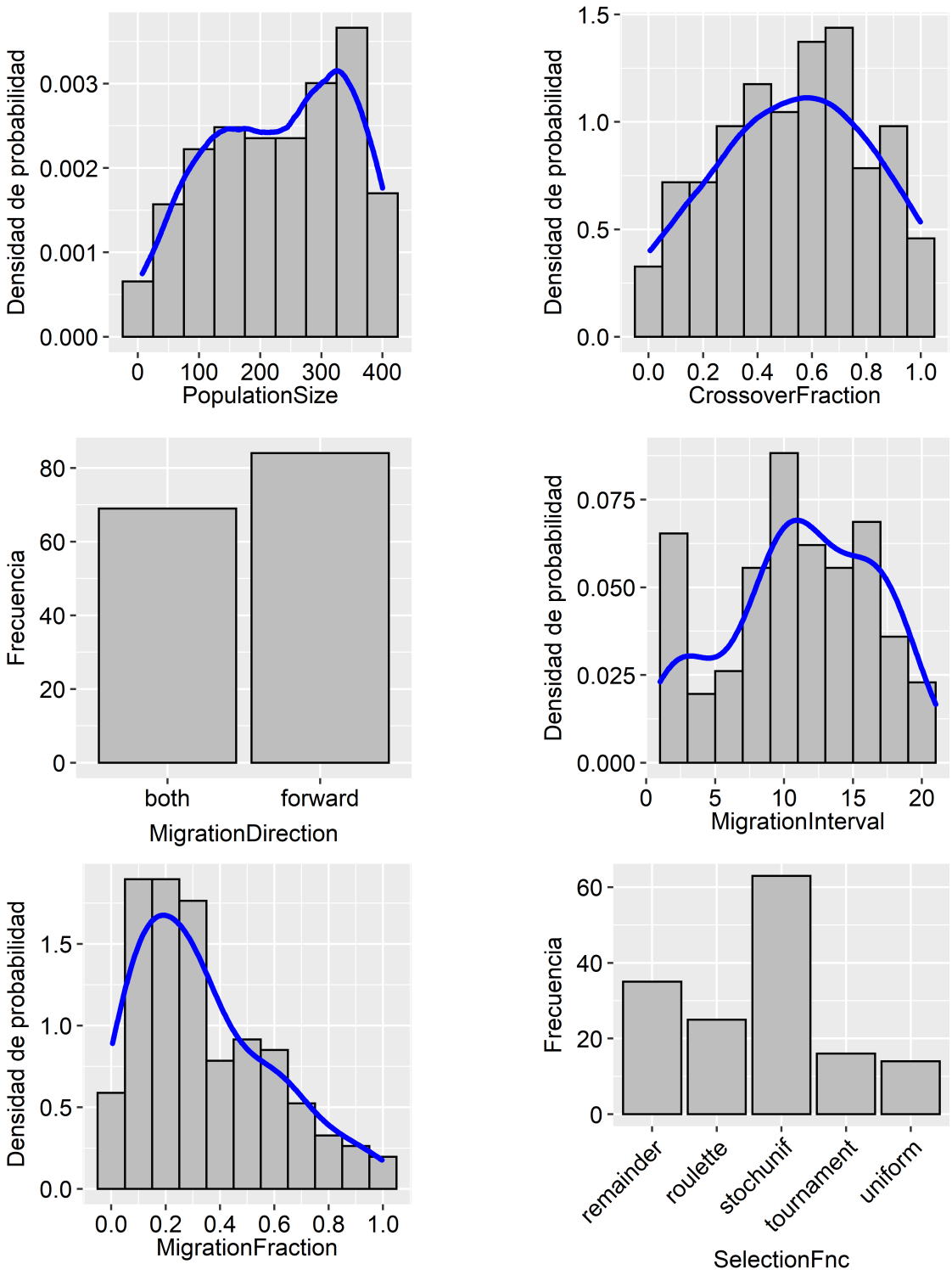


Figura B-32: Diagrama de frecuencia para los parámetros de GA con normalización por valor mínimo para `PopulationSize`, `CrossoverFraction`, `MigrationDirection`, `MigrationInterval`, `MigrationFraction` y `SelectionFnc`.

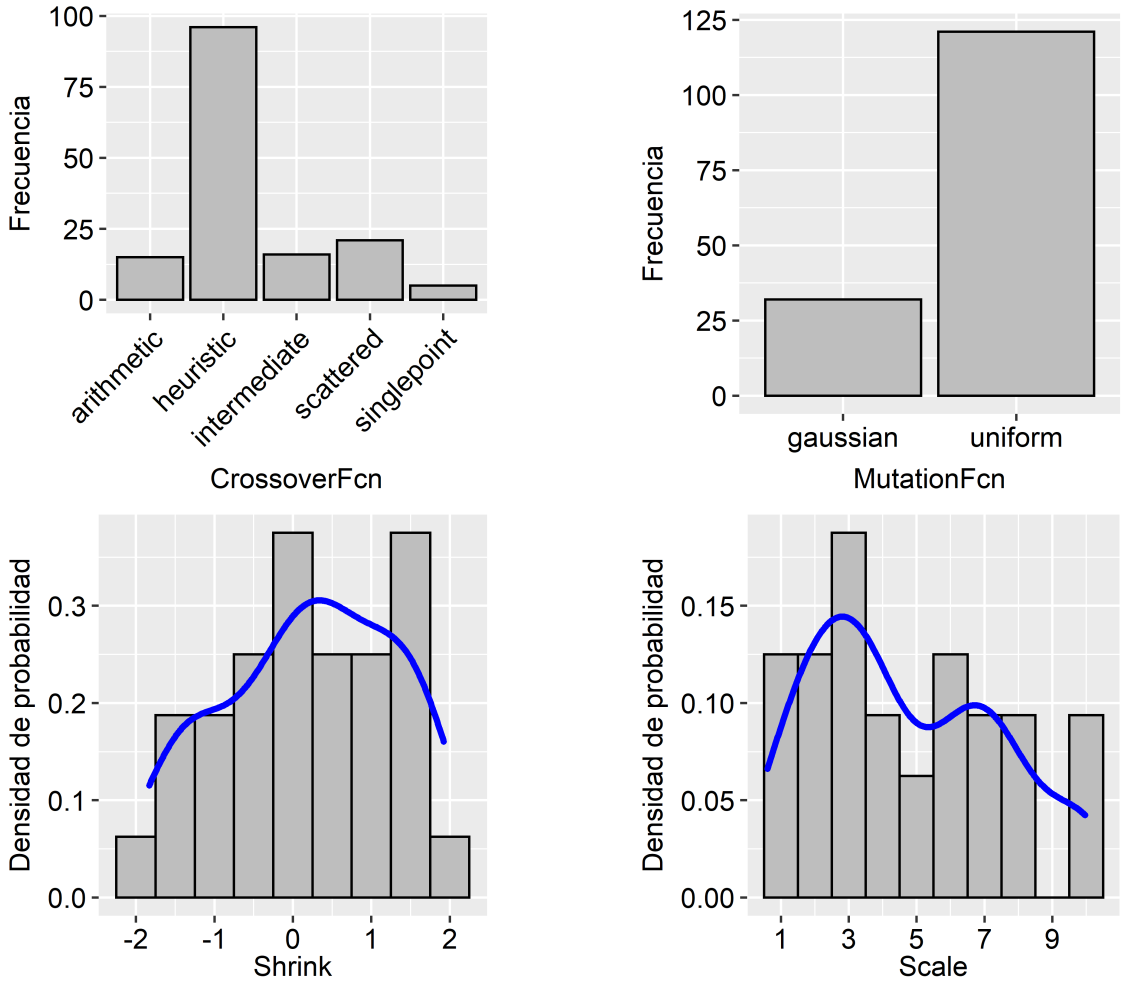


Figura B-33: Diagrama de frecuencia para los parámetros de GA con normalización por valor mínimo para `CrossoverFnc`, `MutationFnc`, `Shrink` y `Scale`.

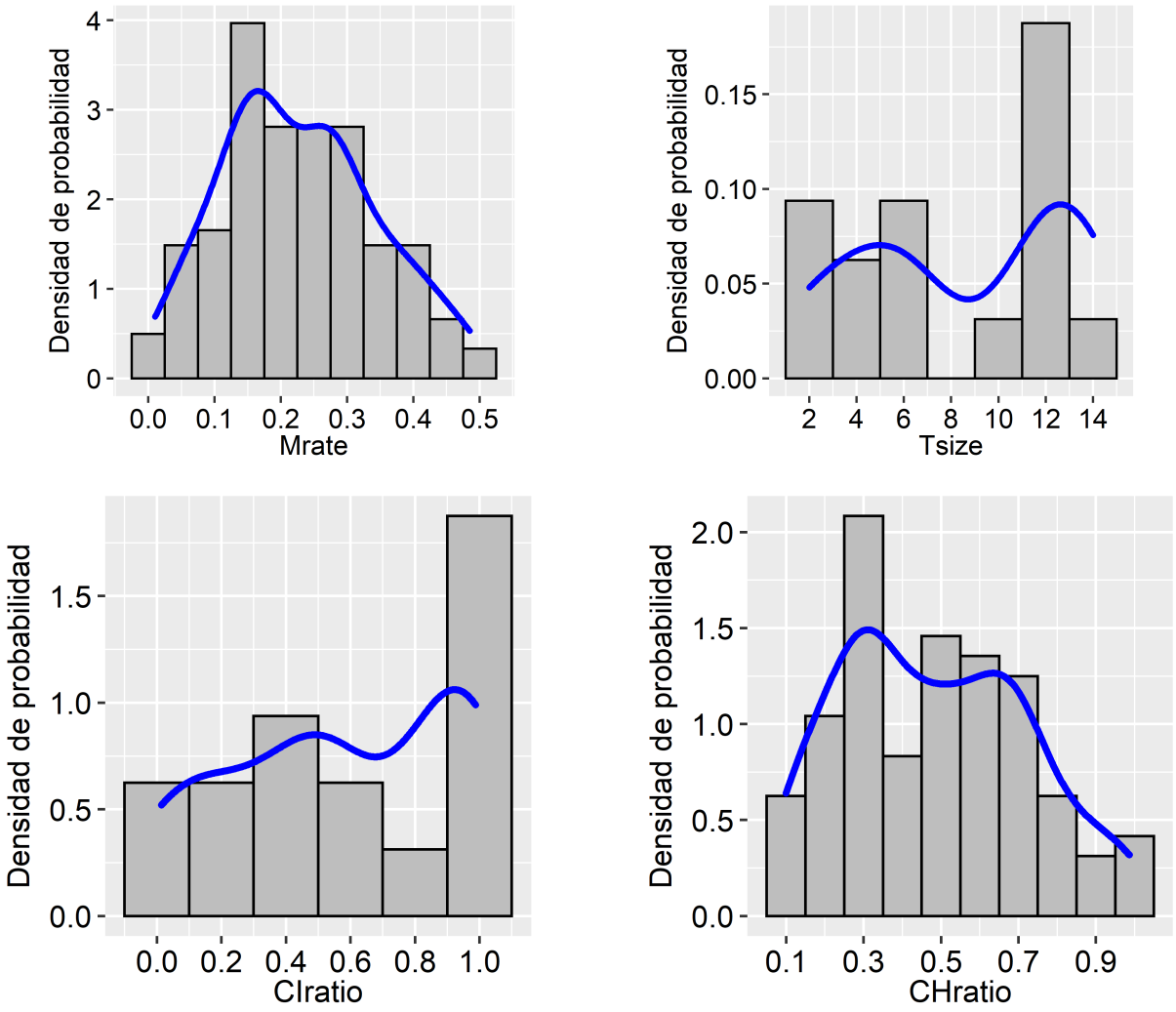


Figura B-34: Diagrama de frecuencia para los parámetros de GA con normalización por valor mínimo para `Mrate`, `Tsize`, `CIratio` y `CHratio`.

B.5. Desempeño de configuraciones sintonizadas

B.5.1. Desempeño de configuraciones con normalización por valor mínimo

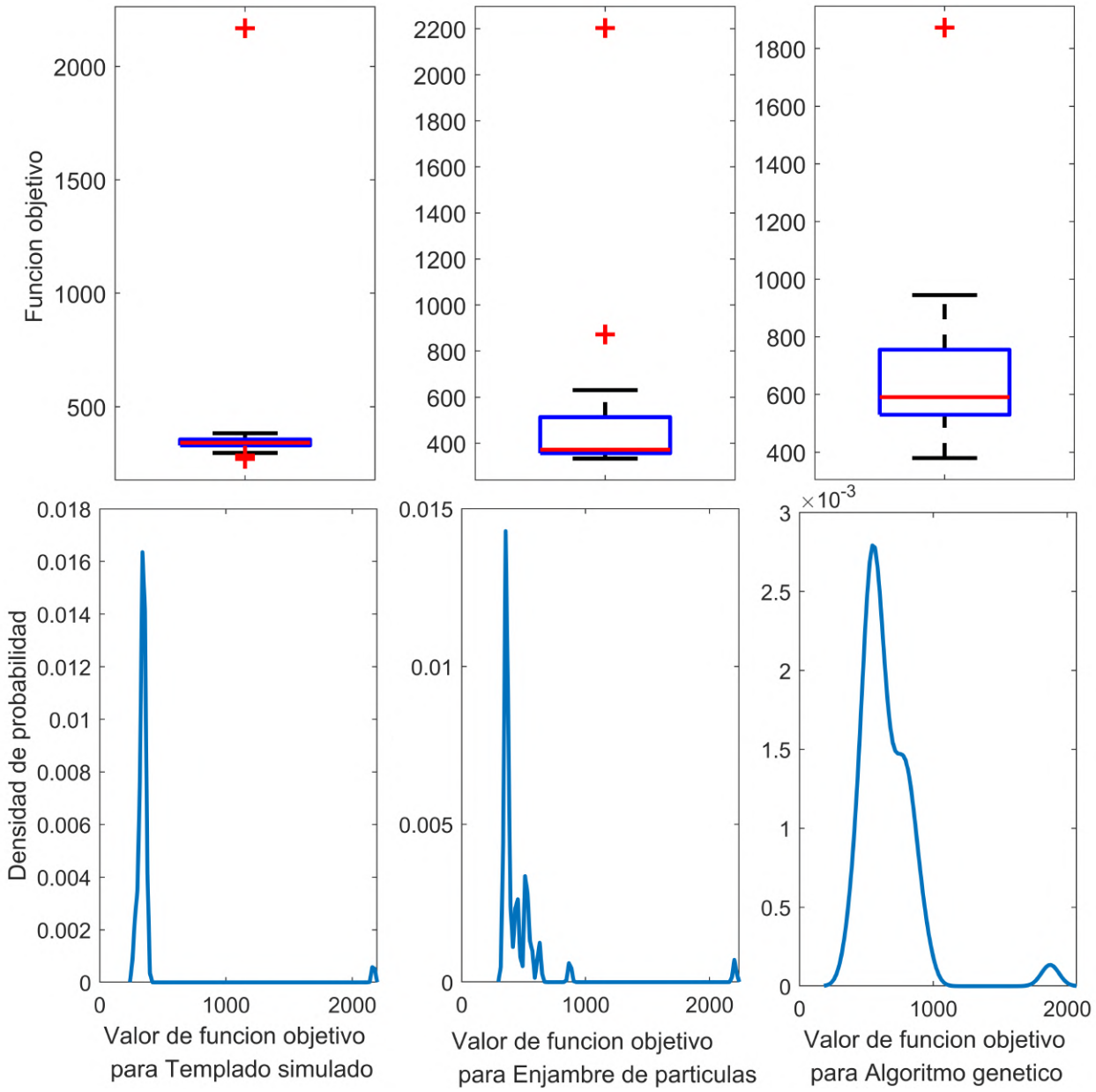


Figura B-35: Desempeño de optimizadores con configuraciones sintonizadas con normalización de valor mínimo.

B.5.2. Desempeño de configuraciones con normalización por valor máximo

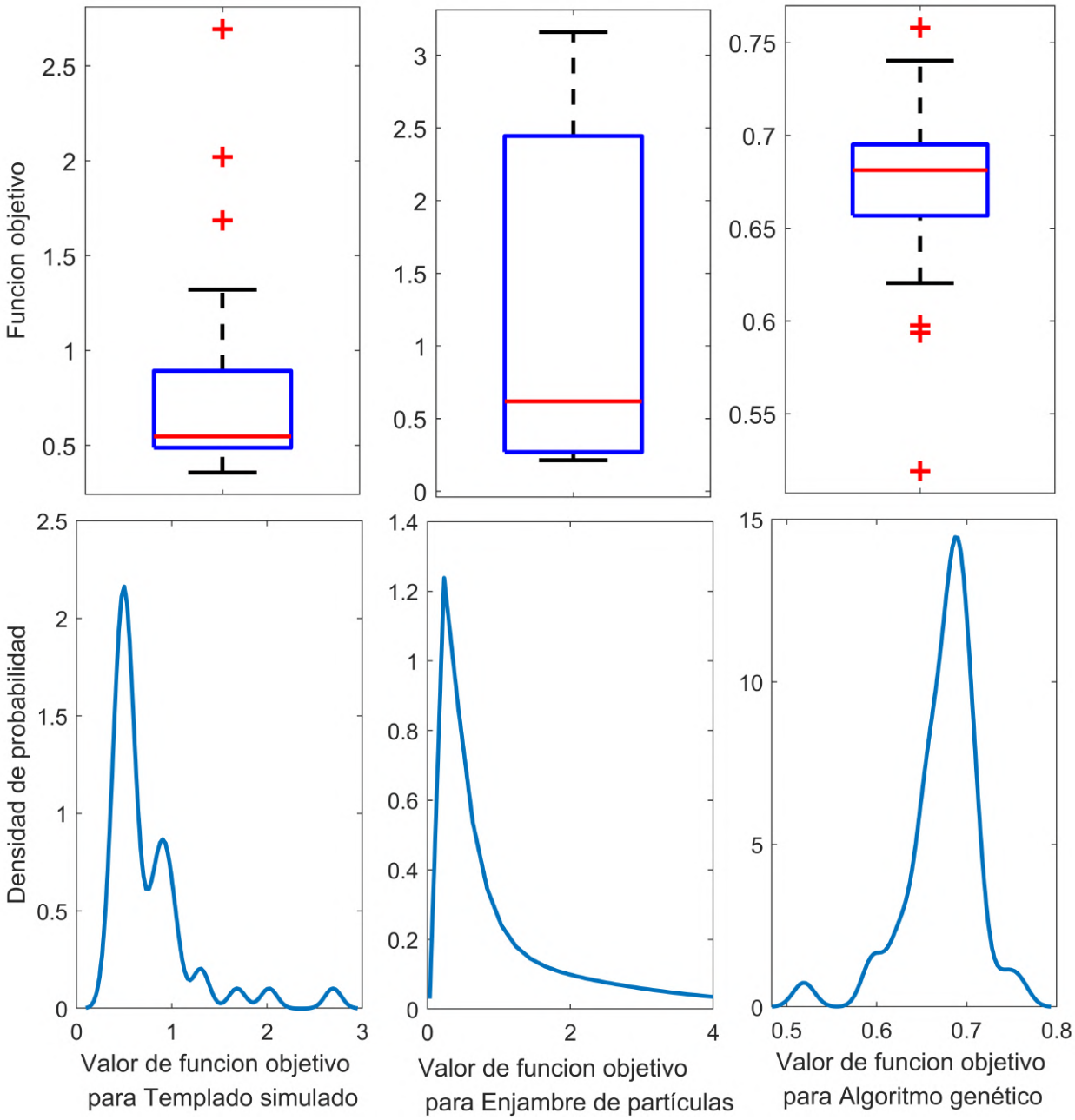


Figura B-36: Desempeño de optimizadores con configuraciones sintonizadas con normalización de valor máximo.

B.5.3. Desempeño de configuraciones con normalización por valor medio

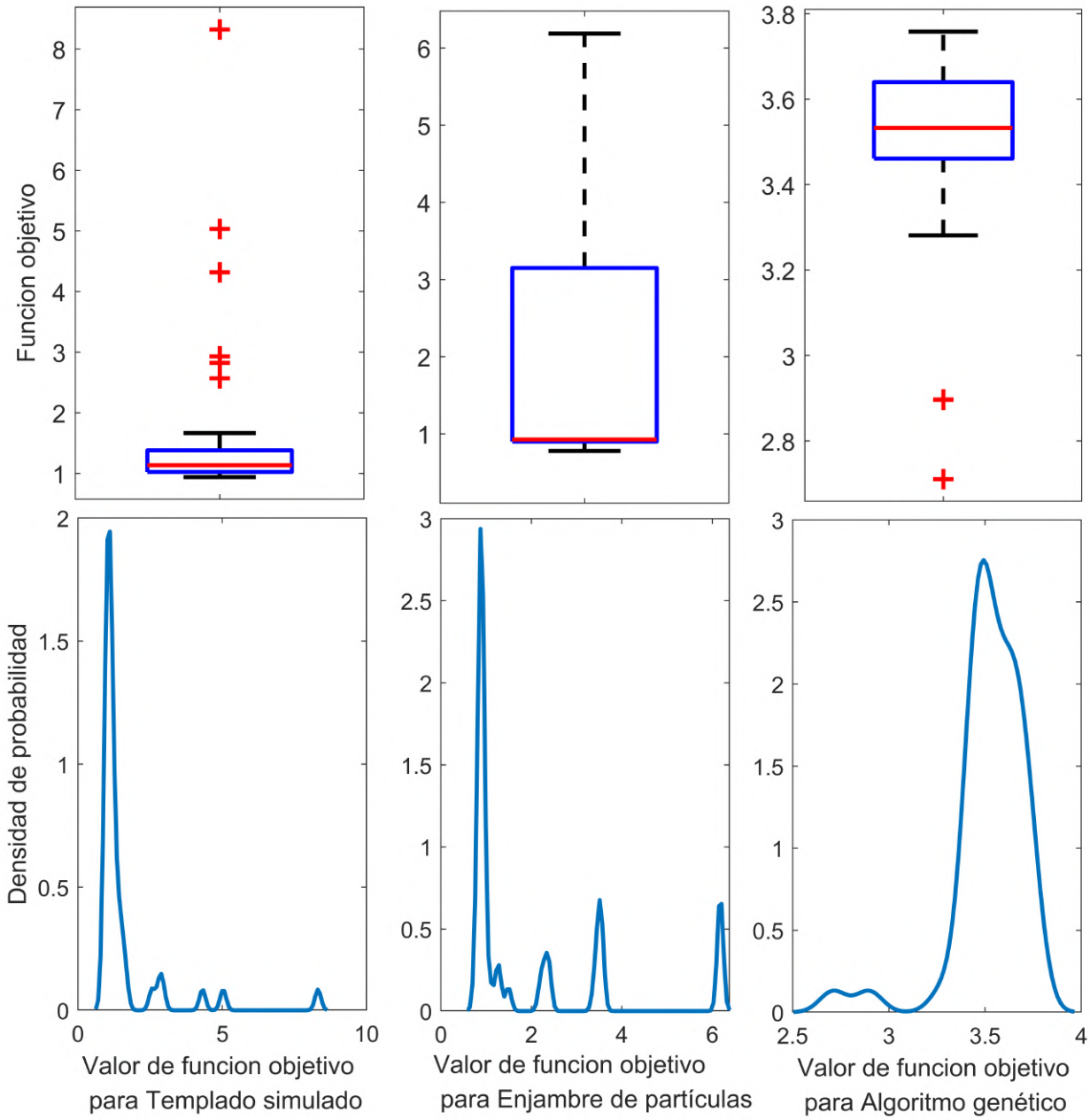


Figura B-37: Desempeño de optimizadores con configuraciones por defecto bajo normalización por valor medio.

B.5.4. Análisis de convergencia teórica del optimizador GA

Tabla B-4: Análisis de convergencia para algoritmo genético bajo diferentes operadores (Schmitt, 2001)

Condición	Caso	Convergencia
Los operadores de mutación y selección son constantes en el tiempo	Entrecruzamiento regular y mutación múltiple	No convergente
	Entrecruzamiento regular, mutación múltiple y selección por templado simulado	Convergente
	Entrecruzamiento regular y selección por función de Boltzmann con aceptación logística	Convergente
La tasa de mutación convergente a un valor positivo	Entrecruzamiento generalizado y mutación simple o múltiple	No convergente
	Entrecruzamiento generalizado, mutación múltiple y selección por ley de potencias	No convergente
Las tasas de mutación y entrecruzamiento varían por intervalos	Entrecruzamiento generalizado, mutación simple o múltiple y selección por templado simulado	Convergente
La tasa de mutación converge a cero	Operador de entrecruzamiento regular constante, mutación múltiple y selección proporcional	No convergente
	Entrecruzamiento generalizado, mutación múltiple y selección proporcional	No convergente
	Selección por ley de potencias	No convergente

Bibliografía

Aarts, E. & Korst, J. (1989). Simulated annealing and Boltzmann machines. New York, NY; John Wiley and Sons Inc.

Birattari, M., Stützle, T., Paquete, L., Varrentrapp, K., et al. (2002). A racing algorithm for configuring metaheuristics. In GECCO, volume 2.

Kennedy, J. & Eberhart, R. (1995). Particle swarm optimization (pso). In Proc. IEEE International Conference on Neural Networks, Perth, Australia (pp. 1942–1948).

Maron, O. & Moore, A. W. (1997). The racing algorithm: Model selection for lazy learners. Artificial Intelligence Review, 11(1-5), 193–225.

Rangaiah, G. P. (2010). Stochastic global optimization: techniques and applications in chemical engineering. World Scientific.

Reynolds, C. W. (1987). Flocks, herds and schools: A distributed behavioral model. Computer Graphics, 21(4), 25–34.

Schmitt, L. M. (2001). Theory of genetic algorithms. Theoretical Computer Science, 259(1-2), 1–61.

Spall, J. C. (2005). Introduction to stochastic search and optimization: estimation, simulation, and control, volume 65. John Wiley & Sons.

Yuan, B. & Gallagher, M. (2004). Statistical racing techniques for improved empirical evaluation of evolutionary algorithms. In International Conference on Parallel Problem Solving from Nature (pp. 172–181).: Springer.

Apéndice C

Estimación y validación de parámetros.

En este apéndice se presenta información complementaria y algunos resultados de los procesos de estimación y validación de parámetros del modelo de fermentación diaúxica de glucosa y xilosa para producción de xilitol.

C.1. Integradores

De forma básica, los algoritmos de integración pueden catalogarse en:

- **Explícitos:** corresponden a métodos numéricos de la forma

$$y^{i+1} = y^i + f(x^i, y^i)h \tag{C-1}$$

  donde y hace referencia al valor del estado en el tiempo i, h es el "tamaño de paso" de la variable independiente x (generalmente el tiempo) y $f(x, y)$ es el modelo matemático. Tienen como característica que el valor a predecir del estado únicamente depende de la variable independiente y de valores pasados de los estados. Por tanto, dicha predicción del estado es *explícita* respecto al lado derecho de la ecuación. Un ejemplo clásico de este tipo de algoritmos son los de tipo "Runge-Kutta".

- **Implícitos:** corresponden a métodos numéricos de la forma

$$y^{i+1} = y^i + f(xi, y^{i+1})h \tag{C-2}$$

  dado que el valor predicho del estado aparece en ambos lados de la ecuación, la predicción del estado es entonces *implícita*. Comúnmente, este tipo de aproximaciones requieren de un proceso iterativo. Un ejemplo clásico es el algoritmo de Heun.

- **De un paso:** pueden ser tanto implícitos como explícitos, se caracterizan porque el valor del estado a predecir solo depende del valor inmediatamente anterior, de la forma

$$y^{i+1} = y^i + f(x^i, y^i)h \tag{C-3}$$

- **De múltiples pasos:** pueden ser tanto implícitos como explícitos, se caracterizan porque el valor del estado a predecir dependerá de un número n de valores previos de dicho estado, de la forma

$$y^{i+1} = y^i + f(x^{i-1}, ..., x^{i-n}, y^{i-1}, ..., y^{i-n})h \tag{C-4}$$

Una posibilidad al momento de integrar sistemas ODE es que un conjunto de parámetros ocasione que el modelo se comporte como *rígido*. Un sistema rígido es aquel que posee estados que cambian rápidamente junto con otros que lo hacen de forma lenta. Desde el punto de vista del método numérico debe considerarse el problema de *estabilidad numérica*, en donde la elección del tamaño de paso dentro del integrador puede llevar a que la solución del sistema EDO sea convergente o divergente. Adicionalmente, los métodos numéricos poseen un error asociado respecto a la solución verdadera del sistema EDO. Estos aspectos afectan directamente la selección del tamaño de paso h, el mismo debe alcanzar una solución convergente con un error aceptable (Chapra et al., 2012). Generalmente, un tamaño de paso menor mejora la precisión y estabilidad de la solución, sin embargo, se aumenta el tiempo de cómputo al aumentar la cantidad de operaciones necesarias en el mismo intervalo de integración para una misma tolerancia de error (Dallas et al., 2017).

C.2. Función objetivo

Algunos ejemplos de diferentes funciones objetivo se presentan a continuación:

- **Función de mínimos cuadrados ponderados** (Englezos & Kalogerakis, 2000):

$$\varphi(\boldsymbol{\theta}) = \sum_{i=1}^{N} \mathbf{e}_i Q_i \mathbf{e}_i^T \tag{C-5}$$

  en donde Q_i corresponde a una matriz de pesos mxm definida por el usuario.

- **Estimador de máxima verosimilitud para una distribución normal** (Englezos & Kalogerakis, 2000):

$$\varphi(\boldsymbol{\theta}) = \sum_{i=1}^{N} \mathbf{e}_i \Sigma_i^{-1} \mathbf{e}_i^T \tag{C-6}$$

en donde bajo suposición de que todos los errores $\mathbf{e}_i$ están normalmente distribuidos, poseen media de cero y una matriz de varianza conocida Σ_i. Esta última corresponde entonces a la inversa de la matriz de covarianza:

$$\Sigma_i = [COV(\mathbf{e}_i)]^{-1} \tag{C-7}$$

la suposición anterior lleva a tres casos en la forma de la matriz Σ_i :

- Caso I: todos los errores son idénticos e independientemente distribuidos con media de cero y varianza σ_e^2:

$$\Sigma_i = \sigma_{\mathbf{e}}^2 I \; ; \varphi(\boldsymbol{\theta}) = \frac{1}{\sigma_{\mathbf{e}}^2} \sum_{i=1}^{N} \mathbf{e}_i \mathbf{e}_i^T \tag{C-8}$$

- Caso II: La varianza de una variable respuesta particular es constante, pero diferentes variables respuesta tienen diferentes varianzas:

$$\Sigma_i = \begin{bmatrix} \sigma_{e1}^2 & 0 & \dots & 0 \\ 0 & \sigma_{e2}^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{em}^2 \end{bmatrix} \; ; V_i = \frac{\sigma_{e1}^2}{\sigma^2} \dots \frac{\sigma_{em}^2}{\sigma^2} \tag{C-9}$$

$$\varphi(\boldsymbol{\theta}) = \sum_{i=1}^{N} \mathbf{e}_i \, diag(V_1^{-1}, \dots, V_m^{-1}) \, e_i^T \tag{C-10}$$

  en este caso V_i corresponde al valor relativo de la varianza para esa variable respuesta.

- Caso III: Cuando la matriz de covarianza es desconocida (situación más común), puede aplicarse el criterio del determinante, el cual consiste en la minimización del determinante de la función objetivo:

$$\varphi(\boldsymbol{\theta}) = det \left(\sum_{i=1}^{N} \mathbf{e}_i \mathbf{e}_i^T \right) \tag{C-11}$$

- **Función objetivo resistente a puntos atípicos** (Aster et al., 2005):

$$\varphi(\boldsymbol{\theta}) = \sum_{i=1}^{N} |\mathbf{e}_i| \tag{C-12}$$

esta forma de la función objetivo, también llamada norma-1 ($||L_1||$), es más **robusta** que la norma-2 (mínimos cuadrados) debido a que los términos no se elevan al cuadrado lo que sobrestimaría el efecto de un *outlier* , así mismo, también tiene posee la cualidad de ser una función de máxima verosimilitud.

- **Prueba de chi-cuadrado** (X^2) (Lillacci & Khammash, 2010):

$$\varphi(\boldsymbol{\theta}) = Prob(X^2, N-p);\ X^2 = \sum_{i=1}^{N} \frac{(\mathbf{e}_i)^2}{y_i} \tag{C-13}$$

esta función objetivo involucra el cálculo de la probabilidad de que los valores calculados por el modelo no sean estadísticamente diferentes a los datos experimentales utilizados en la estimación de parámetros. Esto según la distribución de probabilidad de X^2, para los grados de libertad $N-p$ en donde N corresponde al número de datos experimentales y p al número de parámetros estimados.

- **Optimización regularizada:** este tipo de optimización busca la estabilización de la solución del problema inverso, a través de la adición de un término que incluye información de los parámetros (Aster et al., 2005):

$$\varphi(\boldsymbol{\theta}) = \sum_{i=1}^{N} \mathbf{e}_i \mathbf{e}_i^T + \alpha^2 ||\boldsymbol{\theta}||_2^2 \tag{C-14}$$

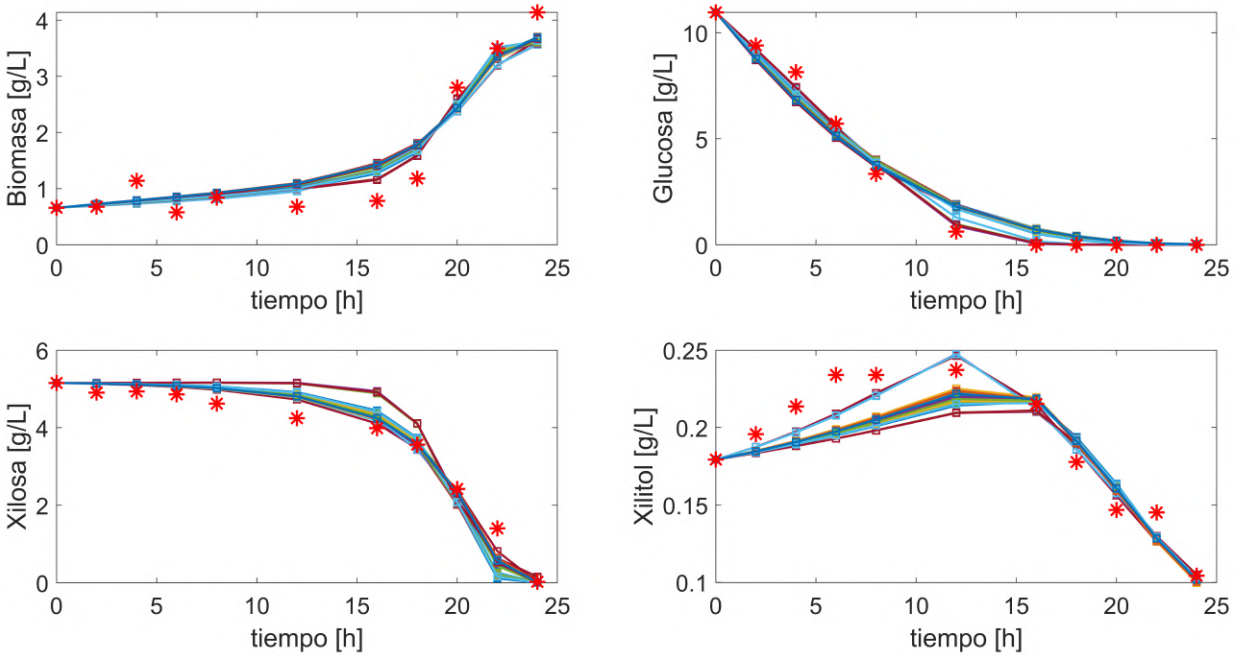
en donde α es el parámetro de regularización. La fórmula anterior corresponde a la regularización de Tikhonov y es particularmente útil en la solución de problemas que sufran de multicolinealidad (interacción entre parámetros), común en modelos matemáticos con alta cantidad de parámetros (Pitt & Banga, 2019).

- **Función de optimización QUAL2Kw:** tomada del framework QUAL2Kw para análisis de calidad de aguas de la U.S. Environmental Protection Agency, la cual busca representar de manera equitativa y robusta todos los estados del modelo al reducir el efecto causado por diferencias en escala. (Pelletier et al., 2006):

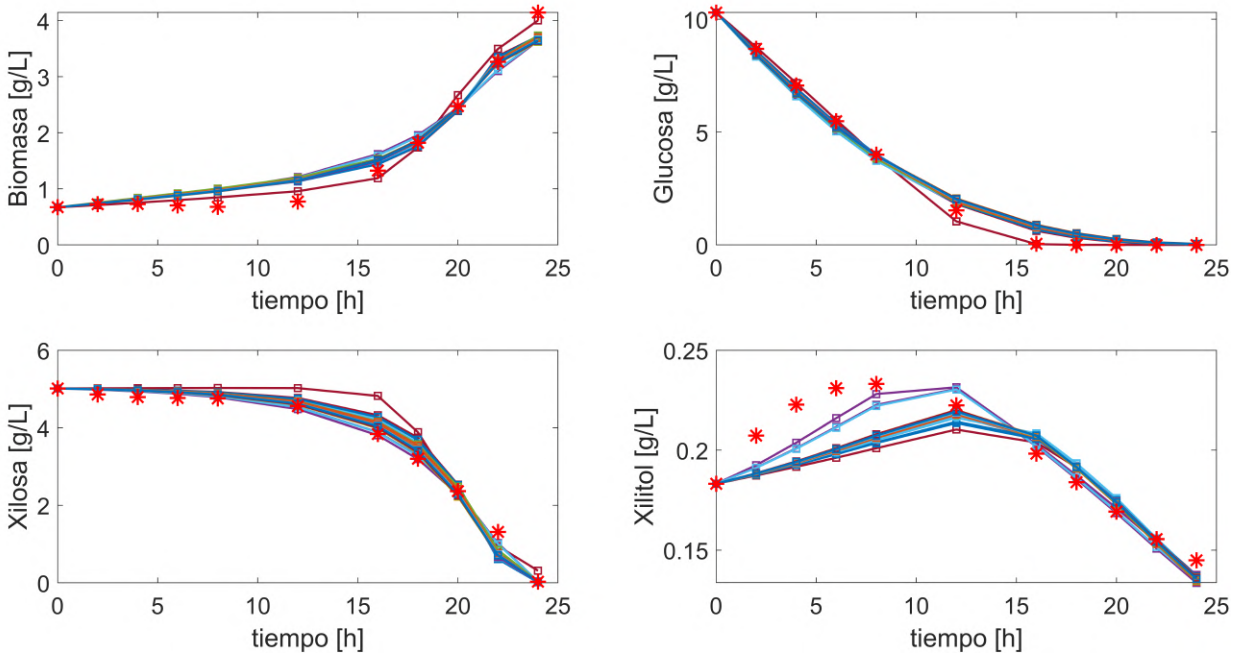
$$f(\boldsymbol{\theta}) = \sum_{i=1}^{q} w_i \sum_{i=1}^{q} \frac{1}{w_i} \frac{\frac{\sum_{j=1}^{N} y_{i,j}^{exp}}{N}}{\sqrt{\frac{\sum_{j=1}^{N} \mathbf{e}_{i,j}}{N}}} \tag{C-15}$$

en donde w_i corresponde a pesos asignados a los estados, q número total de estados y N número total de datos experimentales.

C.3. Tratamiento de datos experimentales



(a) Datos experimentales sin pretratamiento



(b) Datos experimentales con pretratamiento

Figura C-1: Trayectorias del modelo para parámetros estimados con datos experimentales pretratados y no pretratados. (* Datos experimentales, - trayectoria del modelo).

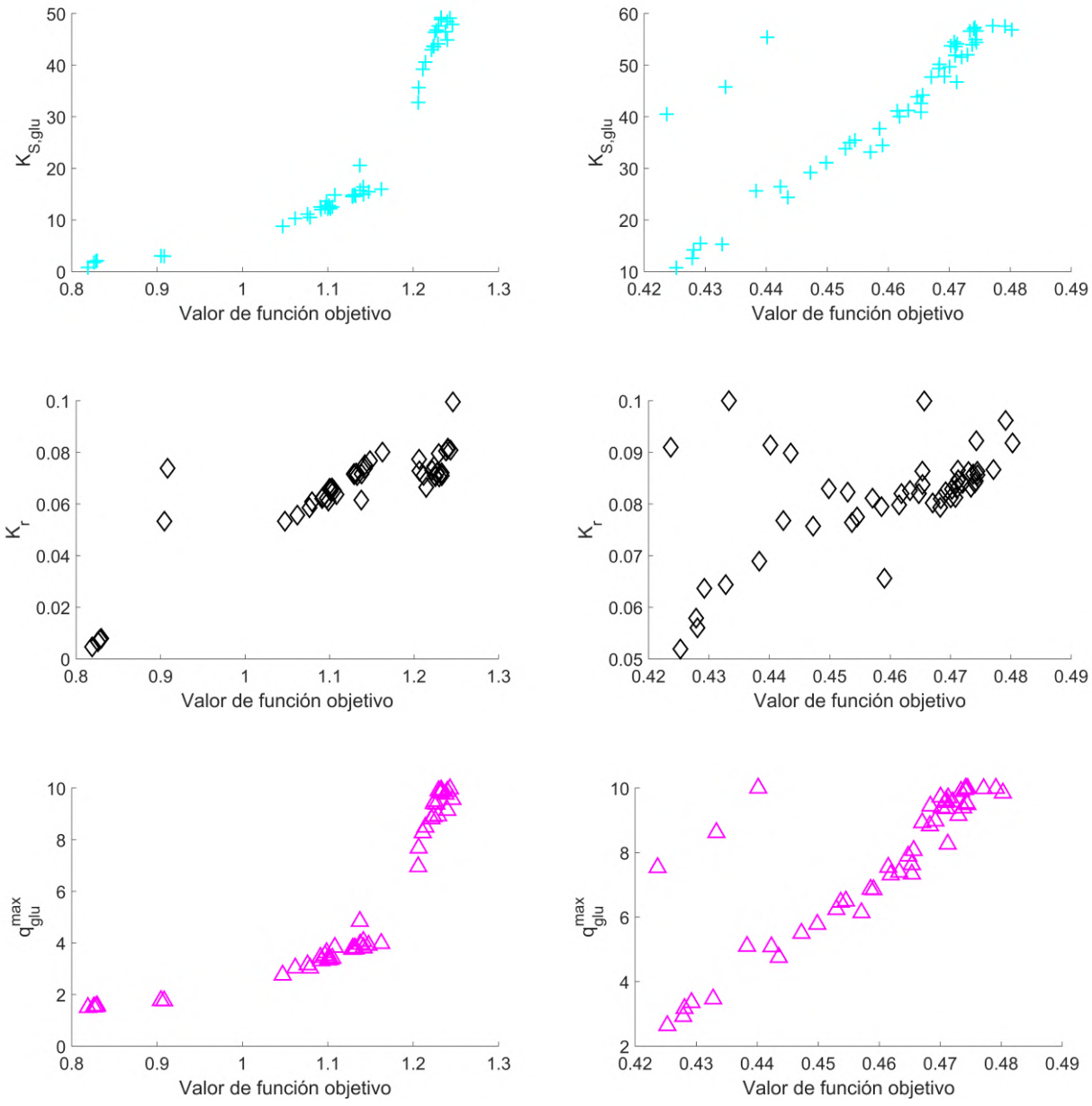


Figura C-2: Valores de parámetros vs valor de función objetivo para los parámetros $K_{S,\,glu}$, K_r y q_{glu}^{max}. Izquierda: datos originales, Derecha: datos pretratados.

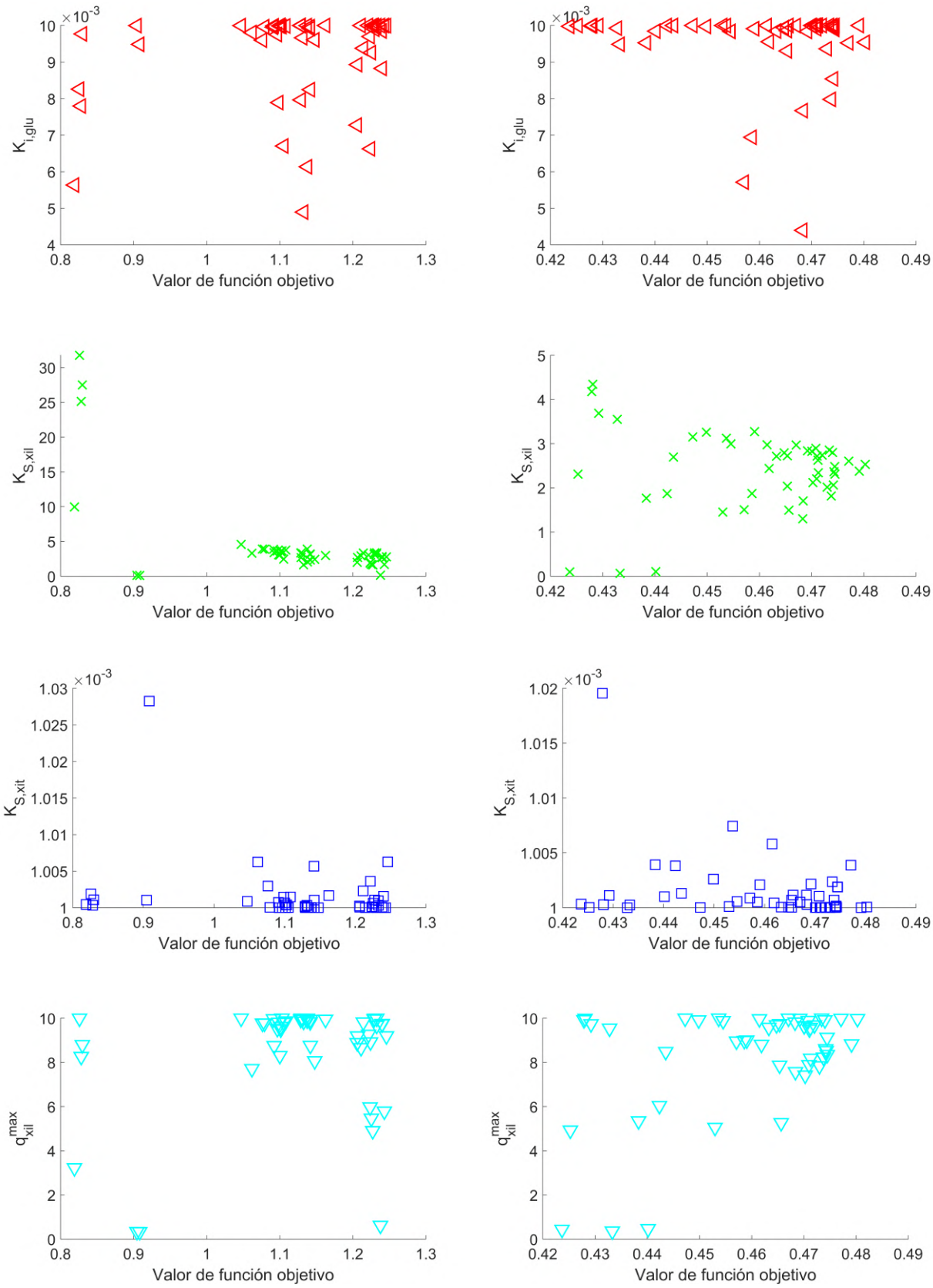


Figura C-3: Valores de parámetros vs valor de función objetivo para los parámetros $K_{i\,glu}$, $K_{S\,xil}$, $K_{S,\,sit}$ y q_{xil}^{max} . Izquierda: datos originales, Derecha: datos pretratados.

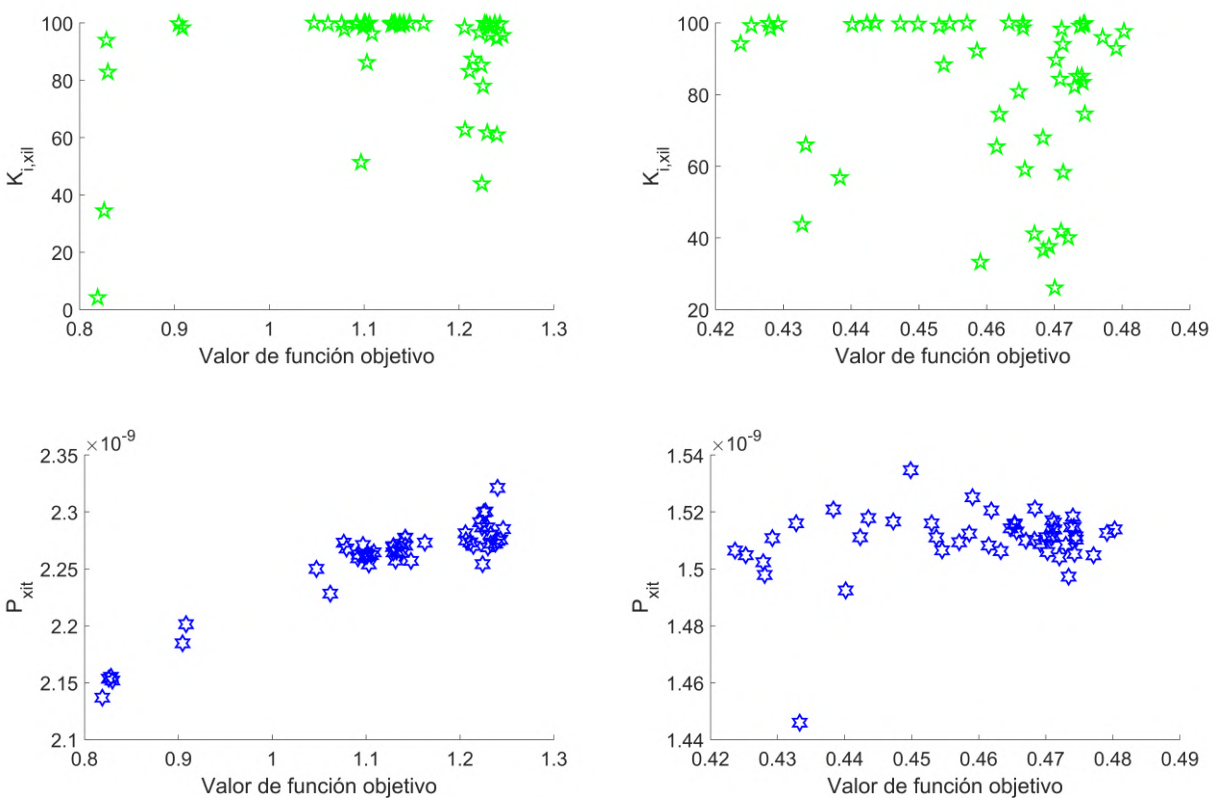


Figura C-4: Valores de parámetros vs valor de función objetivo para los parámetros $K_{i\,xil}$ y P_{xit}. Izquierda: datos originales, Derecha: datos pretratados.

C.4. Identificabilidad estructural

Los resultados de la re-estimación los 11 parámetros con el conjunto de datos 1 aumentado con el comando `spline()` de de 11 a 121 puntos por variable para un total de 484 observaciones experimentales se presentan en las Figuras **C-5** y **C-6** y la Tabla **C-1**.

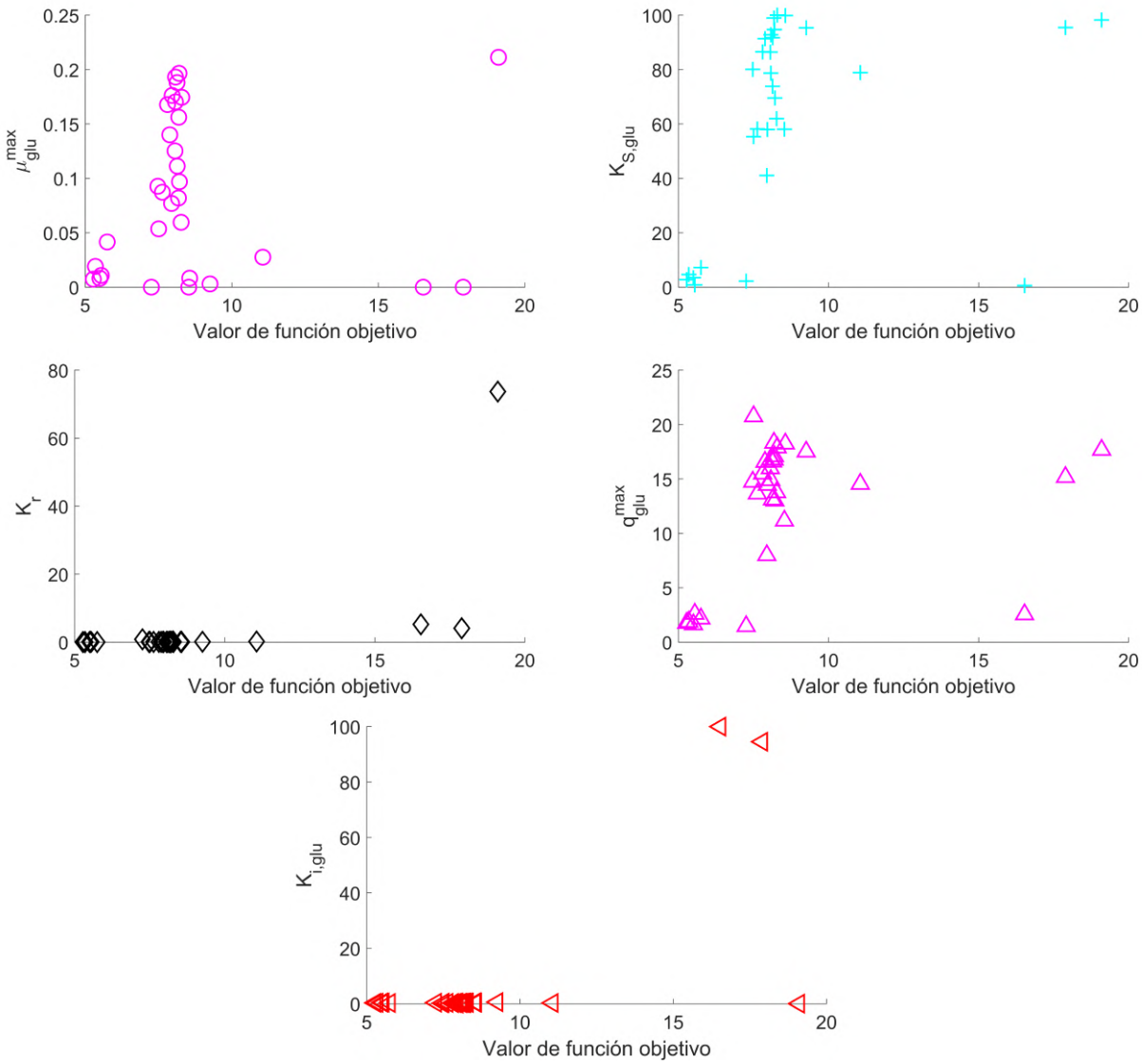


Figura C-5: Estimación de parámetros μ^{max}_{glu}, $K_{S,glu}$, K_r, q^{max}_{glu} y $K_{i,glu}$ con conjunto de datos 1 extendido.

Los parámetros μ^{max}_{glu}, $K_{S,glu}$, K_r, q^{max}_{glu} y $K_{i,glu}$ convergieron a un valor a medida que el optimizador se aproximó al mínimo global y el resto de parámetros se acumuló sobre las cotas o tiene un rango amplio de valores. Lo anterior indica que no se requiere solamente una mayor cantidad de datos experimentales, se quiere nueva información del sistema, la

cual para el caso de modelos matemáticos basados en ecuaciones diferenciales implica experimentos con diferentes condiciones iniciales. Lo anterior se debe a la condición de Lipschitz, la cual establece que la trayectoria de modelos de ecuaciones diferenciales solo depende de la estructura, parámetros y condiciones iniciales del modelo y es única dentro de un intervalo de la variable independiente (Teschl, 2012). Por tanto, condiciones iniciales diferentes aportan nuevas trayectorias y con ello mayor información.

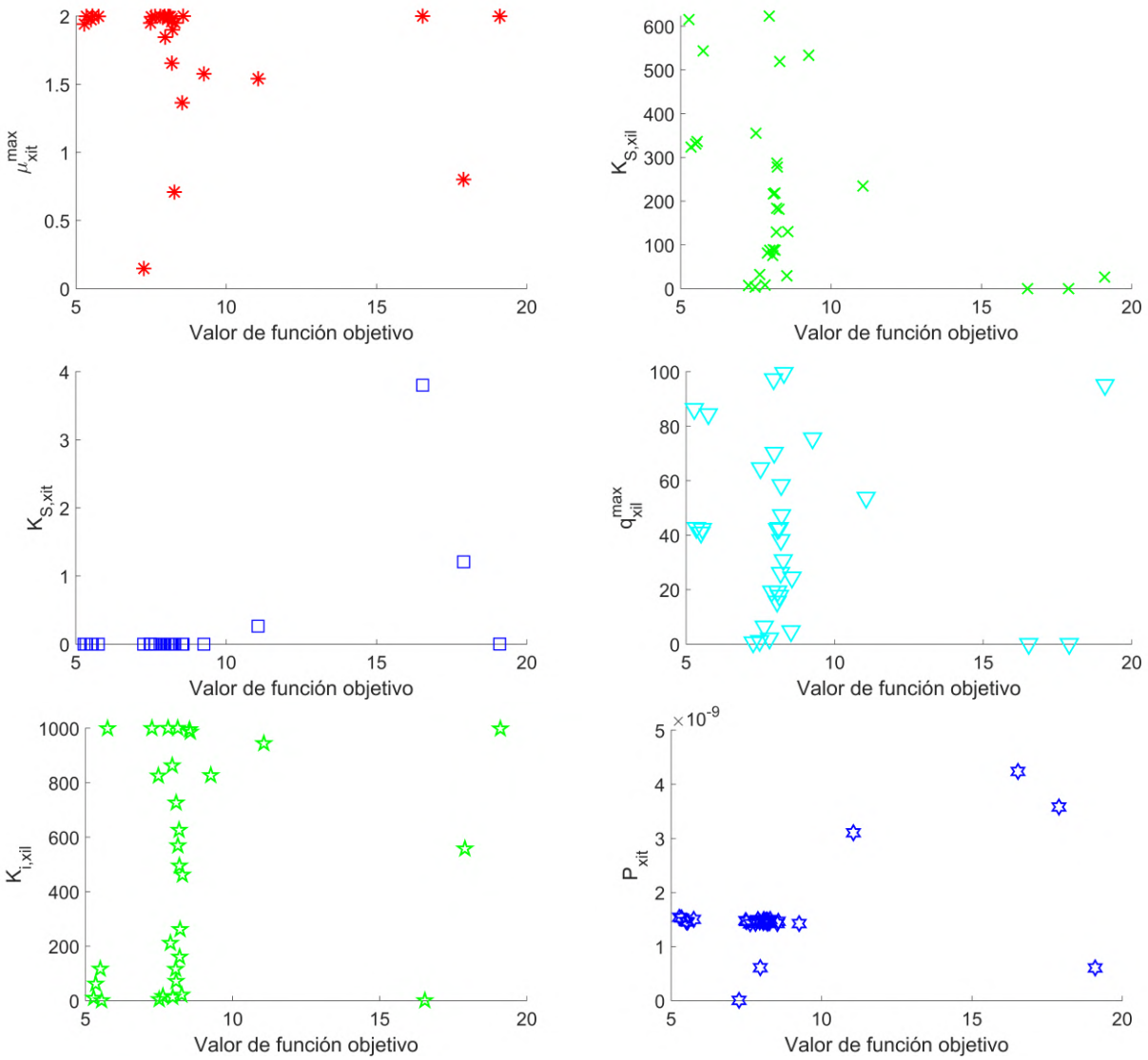


Figura C-6: Estimación de parámetros μ^{max}_{xit}, $K_{S,xil}$, $K_{S,xit}$, q^{max}_{glu}, $K_{i,xil}$ y P_{xit} con conjunto de datos 1 extendido.

Tabla C-1: Parámetros identificados conjunto de datos 1 aumentado con `spline()`.

Parámetro	Valor	Parámetro	Valor	Parámetro	Valor
μ_{glu}^{max}	1.082x10^{-2}	$K_{S,\,glu}$	3.4696	q_{glu}^{max}	1.869
$K_{i,\,glu}$	0.329	K_r	4.63x10^{-2}	μ_{xit}^{max}	1.9988
$K_{S,\,xil}$	336.402	$K_{S,\,xit}$	8.4857x10^{-5}	q_{xil}^{max}	42.6902
$K_{i,\,xil}$	61.34	P_{xit}	1.5023x10^{-9}	-	-

C.5. Correlación de parámetros

Tabla C-2: Parámetros correlacionados.

Grupo	Parámetros	Tipo de interacción
1	q_{xil}^{max}, $K_{S,\,glu}$	Inversa
2	μ_{xit}^{max}, K_r	Inversa
3	μ_{xit}^{max}, $K_{S,\,xit}$	Inversa
4	$K_{S,\,xil}$, $K_{i,\,glu}$	Directa
5	$K_{S,\,xit}$, K_r	Directa
6	q_{glu}^{max}, $K_{i,\,xil}$	Inversa

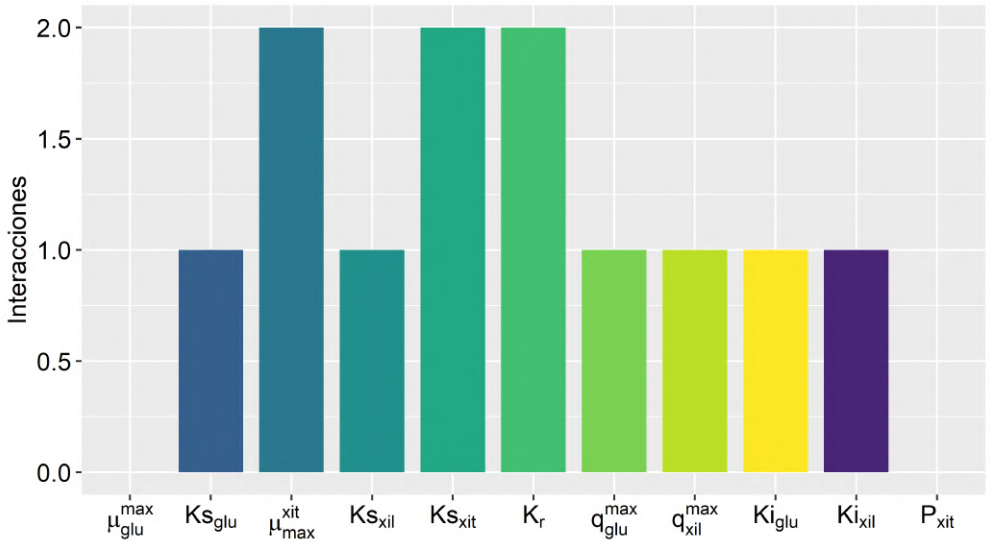


Figura C-7: Frecuencias de interacciones por parámetro.

C.6. Indicadores de ajuste

Tabla C-3: Coeficiente de determinación R^2 para estimación.

Conjunto de datos	Biomasa		Glucosa		Xilosa		Xilitol	
	Lineal	No lineal	Lineal	No lineal	Lineal	No lineal	Lineal	No lineal
				Estimación				
1	-0.146	0.225	0.977	0.975	0.376	0.316	-828.7	0.490
2	-1.283	0.363	NA	0.500	-1.186	0.392	0.567	0.731
3	0.340	0.527	NA	0.500	0.985	0.987	0.905	0.926
4	-13.69	0.520	NA	0.500	0.902	0.902	0.948	0.952
5	-0.086	0.735	NA	0.500	0.983	0.983	0.991	0.992
6	0.817	0.847	NA	0.500	0.983	0.982	0.914	0.920
7	0.674	0.743	NA	0.500	0.997	0.998	0.991	0.991
8	-8.930	0.500	NA	0.500	0.871	0.915	0.980	0.980
9	-3.009	0.473	NA	0.500	0.742	0.862	0.915	0.904
Promedio	-2.813	0.548	0.977	0.975	0.628	0.815	0.901	0.8768
				Validación				
10	-1.836	0.397	NA	0.500	-0.385	0.381	-3.070	0.349
11	-11.24	0.424	NA	0.500	0.674	0.833	0.857	0.909
12	-0.244	0.457	NA	0.500	0.273	0.735	0.896	0.904
13	-7.418	0.513	NA	0.500	0.423	0.730	0.768	0.764
14	0.622	0.761	NA	0.500	0.125	0.691	0.385	0.682
15	-12.80	0.471	NA	0.500	-0.621	0.616	0.628	0.653
16	-2.606	0.353	NA	0.500	0.365	0.603	0.333	0.488
17	-0.802	0.333	NA	0.500	0.415	0.437	-14.93	0.430
18	-1.387	0.429	NA	0.500	0.863	0.880	-72.82	0.464
19	-18.96	0.474	NA	0.500	0.388	0.719	0.536	0.753
20	-108.1	0.479	NA	0.500	-0.080	0.642	0.728	0.667
21	0.103	0.776	NA	0.500	0.142	0.689	0.658	0.699
22	-8.855	0.572	NA	0.500	-0.050	0.629	0.744	0.708
Promedio	-13.35	0.495	NA	0.500	0.138	0.660	0.653	0.651

C.7. Análisis de autocorrelación

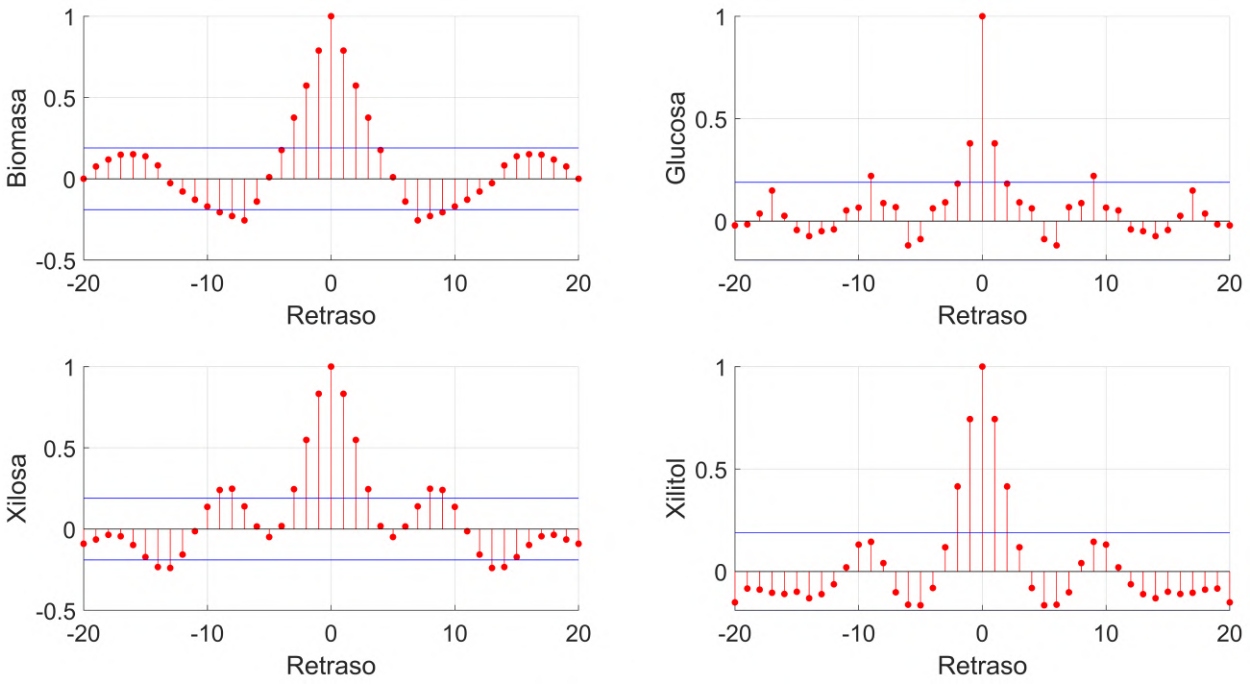


Figura C-8: Análisis de autocorrelación para los residuales de los conjuntos de datos 10 a 22. (— intervalos de confianza, $\alpha = 95\,\%$).

C.8. Método de Sobol

Este procedimiento se basa en el teorema de descomposición de la varianza propuesta por Sobol (Sobol, 1993). El teorema establece que si todas los entradas (parámetros) son independientes, la *varianza total* V *de la respuesta del modelo* puede ser descompuesta en la siguiente serie finita:

$$V = Var[f(\theta)] = \sum_{i=1}^{p} V_i + \sum_{i=1}^{p}\sum_{j>i}^{p} V_{i,j} + \cdots + V_{1,2,\ldots,p} \tag{C-16}$$

El primer término hace referencia a la suma de todas las varianzas parciales individuales de cada entrada, el segundo término consiste en las varianzas debido a la incertidumbre conjunta (interacción) en cada par de entradas. El termino final es la varianza conjunta de todos las entradas. El primer y segundo término de la varianza total están dados por:

$$V_i = Var[E(x|\theta_i)],\ i = 1, 2, ..., p \tag{C-17}$$

$$V_{i,j} = Var[E(x|\theta_i, \theta_j)] - V_i - V_j \tag{C-18}$$

cuando $V_{i,j}$ es diferente de cero, los parámetros θ_i y θ_j son *interactuantes*. En esta caso la varianza total tiene una mayor contribución debido a los dos factores que la suma $V_i + V_j$ por sí misma.

Debido a que la descomposición de la varianza total tiene 2^p términos que requieren ser estimados, su cálculo directo es impráctico en la realidad. Por razón se propusieron los índices de sensibilidad global, definidos como:

$$S_i = \frac{V_i}{V} \tag{C-19}$$

$$S_i^T = \frac{E[Var(x|\boldsymbol{\theta_{-i}})]}{V} = \frac{V - Var[E(x|\boldsymbol{\theta_{-i}})]}{V} = 1 - \frac{V_{-i}}{V} \tag{C-20}$$

en donde V_{-i} indica la varianza de la salida debido a todos los parámetros excepto θ_i, S_i es el *índice de sensibilidad global de primer orden* o índice de Sobol de primer orden que mide el efecto *principal* de la incertidumbre del parámetro θ_i en la varianza de la respuesta. S_i^T es el *índice de sensibilidad de orden total* y cuantifica el efecto *total* de la incertidumbre del parámetro θ_i en la respuesta del modelo.

Si el modelo es no lineal $S_i < S_i^T$ y $\sum S_i^T > 1$ debido a la interacción entre los parámetros θ_i y θ_j es tenida en cuenta tanto en S_i^T como en S_j^T.

Para calcular entonces los índices de sensibilidad global se utiliza el algoritmo propuesto por Saltelli (Saltelli, 2002; Saltelli et al., 2008, 2010; Plischke et al., 2013; Ye & Hill, 2017):

- 1. Generación de dos matrices independientes A y B que contienen un muestreo de los parámetros $\theta \in \Theta$. Posteriormente se genera una matriz C_i mediante el reemplazo la columna i de la matriz A con la columna i de la matriz B:

$$A = \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1p} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2p} \\ \vdots & \vdots & \vdots & \ddots \\ \theta_{N1} & \theta_{N2} & \cdots & \theta_{Np} \end{bmatrix}, B = \begin{bmatrix} \theta'_{11} & \theta'_{12} & \cdots & \theta'_{1p} \\ \theta'_{21} & \theta'_{22} & \cdots & \theta'_{2p} \\ \vdots & \vdots & \vdots & \ddots \\ \theta'_{N1} & \theta'_{N2} & \cdots & \theta'_{Np} \end{bmatrix}$$
$$C = \begin{bmatrix} \theta'_{11} & \theta'_{12} & \cdots & \theta_{1i} & \cdots & \theta'_{1p} \\ \theta'_{21} & \theta'_{22} & \cdots & \theta_{2i} & \cdots & \theta'_{2p} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta'_{N1} & \theta'_{N2} & \cdots & \theta_{Ni} & \cdots & \theta'_{Np} \end{bmatrix} \tag{C-21}$$

en donde N es el número de muestras. Una forma eficiente de generar las matrices A y B es a través del método de muestreo por hipercubo latino o Latin Hypercube Samplig (LHS) (Sheikholeslami & Razavi, 2017). Este tipo de muestreo garantiza una distribución uniforme de las muestras en el espació paramétrico, lo que permite obtener mayor cantidad de información de la sensibilidad de los parámetros con el uso de una menor cantidad de muestras. Adicionalmente, Saltelli y colaboradores propusieron una modificación a la matriz C_i para que esta sea útil tanto para el cálculo de S_T como S_i mediante *muestreo radial* (Saltelli et al., 2010). Este tipo de muestreo fue diseñado específicamente para ser utilizado con estimadores de Monte Carlo, debido a que permite un "paso" en la dirección del parámetro θ_i de la siguiente manera:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \cdots & a_{1,p} \\ b_{1,1} & a_{1,2} & a_{1,3} & \cdots & a_{1,p} \\ a_{1,1} & b_{1,2} & a_{1,3} & \cdots & a_{1,p} \\ a_{1,1} & a_{1,2} & b_{1,3} & \cdots & a_{1,p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{1,1} & a_{1,2} & c_{1,3} & \cdots & b_{1,p} \\ b_{1,1} & b_{1,2} & b_{1,3} & \cdots & b_{1,p} \end{bmatrix} \tag{C-22}$$

En donde cada fila corresponde a una muestra de p parámetros y la primera fila es tomada de la matriz A, segunda fila de la matriz C_1, tercera fila de C_2 y así sucesivamente hasta la fila C_p. Finalmente, para se adiciona una fila de la matriz B, lo que genera un diseño en bloques de $p + 2$ muestras. El número total de muestras para realizar el cálculo de los índices de sensibilidad es $N = k(p + 2)$, en donde k es el número de bloques a utilizar. Este diseño experimental tiene otra ventaja, la cual consiste en ser un diseño internamente balanceado, lo que implica que cada bloque posee la misma

cantidad de efectos por parámetro. Lo anterior permite utilizar un número arbitrario k de bloques.

- 2. Se usan las matrices A, B y C_i para calcular las salidas del modelo, en donde se obtienen:

$$x_A = \{x_{A,1}, x_{A,2}, ..., x_{A,N}\}\,,\ x_B = \{x_{B,1}, x_{B,2}, ..., x_{B,N}\}\,, \tag{C-23}$$

$$x_{C_i} = \{x_{C,1}, x_{C,2}, ..., x_{C_i,N}\} \tag{C-24}$$

- 3. Calcular los estadísticos de las muestras

$$E = \frac{1}{N}\sum_{j=1}^{N} f(A_j) \tag{C-25}$$

$$V = \frac{1}{N}\sum_{j=1}^{N} f^2(A_j) - E^2 \tag{C-26}$$

$$V_i = Var[E(x|\theta)] = \frac{1}{N}\sum_{j=1}^{N} f(B_j)\left(f(C_{i,j}) - f(A_j)\right) \tag{C-27}$$

$$V_{-i} = Var[E(x|\boldsymbol{\theta_{-i}})] = \frac{1}{2N}\sum_{j=1}^{N}(f(A_j) - f(C_{i,j}))^2 \tag{C-28}$$

- Usar los estadísticos del paso anterior para calcular S_i y S_i^T

Los modelos matemáticos de procesos biotecnológicos comúnmente son expresado en forma de sistemas de ecuaciones. Dado que en general estos modelos poseen varios estados y, en el caso de ser compuestos por ecuaciones diferenciales deben ser resueltos para un intervalo de tiempo, los índices S_i y S_i^T deben ser calculados por estado y punto de tiempo (Weber et al., 2018).

C.9. Índices de Sobol

C.9.1. Muestreo con distribución uniforme

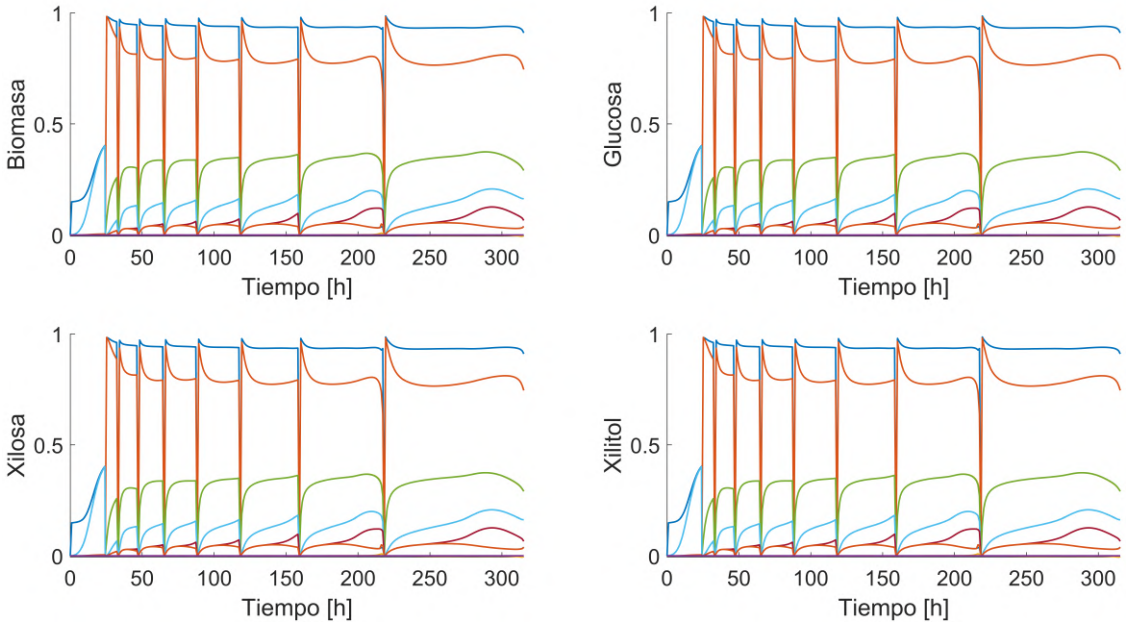


Figura C-9: Perfiles de índices individuales de Sobol por estado para el modelo matemático de fermentación diaúxica para producción de xilitol con distribución uniforme.

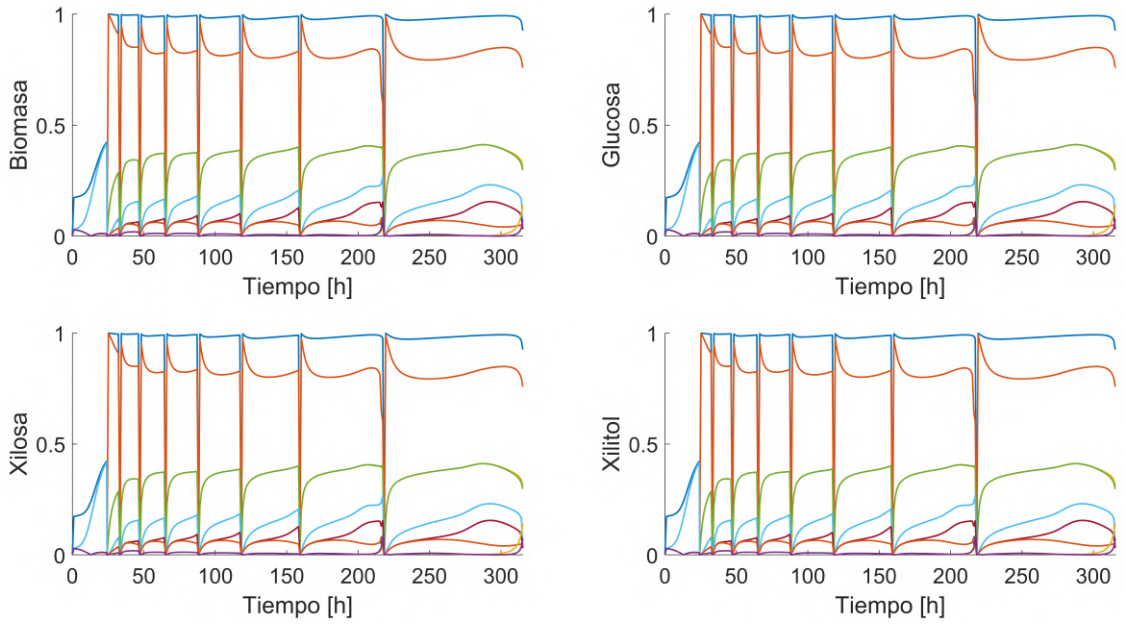


Figura C-10: Perfiles de índices totales de Sobol por estado para el modelo matemático de fermentación diaúxica para producción de xilitol con distribución uniforme.

C.9.2. Muestreo con distribución normal y matriz de covarianza

La Figura **C-11** presenta los perfiles de índices de sensibilidad individuales calculados con matriz de muestreo realizada con distribución normal y matriz de covarianza de los parámetros. Los perfiles de sensibilidad son idénticos para los cuatro estados del modelo, por tanto, la Figura **C-12** presenta los perfiles de índices de sensibilidad individual por parámetro. Bajo este tipo de muestreo, no se observa sensibilidad del modelo hacia ningún parámetros en el conjunto de datos 1. Para los conjuntos de datos 2 a 9, los parámetros con mayor sensibilidad son μ_{glu}^{max} y $K_{S,glu}$ cuyos perfiles se superponen. Posteriormente, se tienen μ_{xit}^{max}, $K_{S,xil}$ y $K_{S,xit}$ con igual sensibilidad. El modelo sería insensible al resto de parámetros. En comparación con la Figura **5-10** y dado que los parámetros están altamente correlacionados, el muestreo realizado con distribución normal y matriz de covarianza aumentaría de manera no deseada el efecto de la correlación en los índices de sensibilidad individuales.

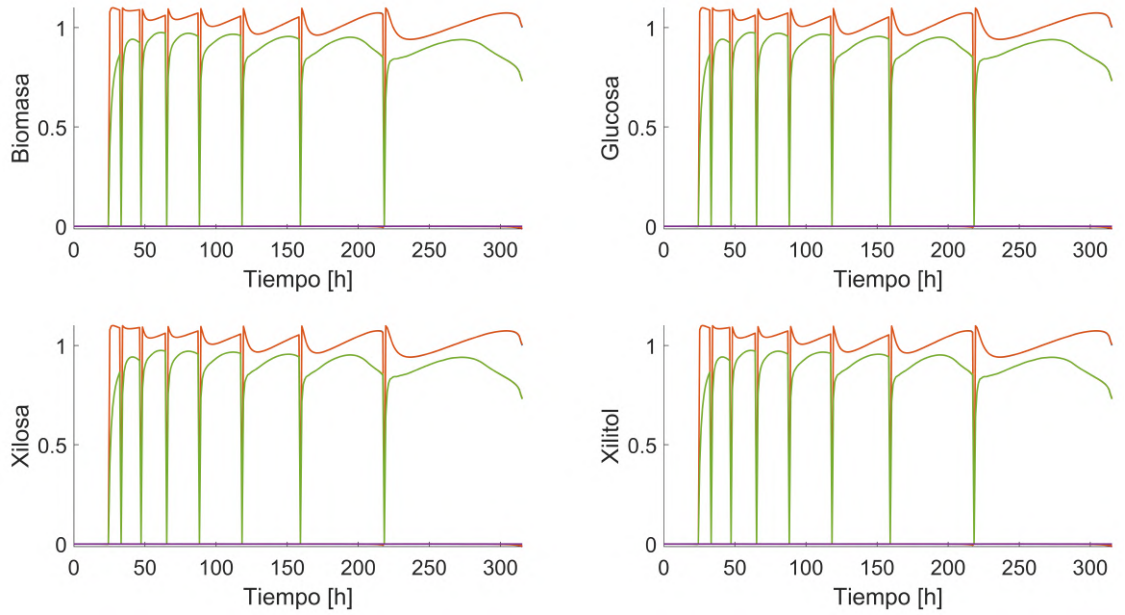


Figura C-11: Perfiles de índices totales de Sobol por estado para el modelo matemático de fermentación diaúxica para producción de xilitol con distribución normal.

Por otra parte, la Figura **C-13** presenta los índices de sensibilidad totales para el modelo de caso de estudio con muestro por distribución normal y matriz de covarianza. Nuevamente, los perfiles de índices de sensibilidad son idénticos para los cuatros estados, por tanto, la Figura **C-14** presenta los perfiles de índices de sensibilidad totales detallados por parámetros para el modelo de caso de estudio con muestro por distribución normal y matriz de covarianza. Para el conjunto de datos 1, los índices S_{T_i} son cercanos a la unidad para todos los parámetros y para los conjuntos 2 a 9 el efecto de interacción es mucho más evidente, llegando a superar el valor de 2 unidades de manera consistente. Adicionalmente, en estos conjuntos de datos

se forman tres grupos de parámetros con la misma sensibilidad, en orden descendente se tienen: μ_{glu}^{max} y $K_{S,glu}$, seguido de μ_{xit}^{max}, $K_{S,xit}$ y $K_{S,xil}$ y por último, K_r, q_{glu}^{max}, q_{xil}^{max}, $K_{i,glu}$, $K_{i,xil}$ y P_{xit}. En vista de estos resultados, el muestreo por distribución uniforme con matriz de covarianza no es recomendable en caso de parámetros altamente correlacionados.

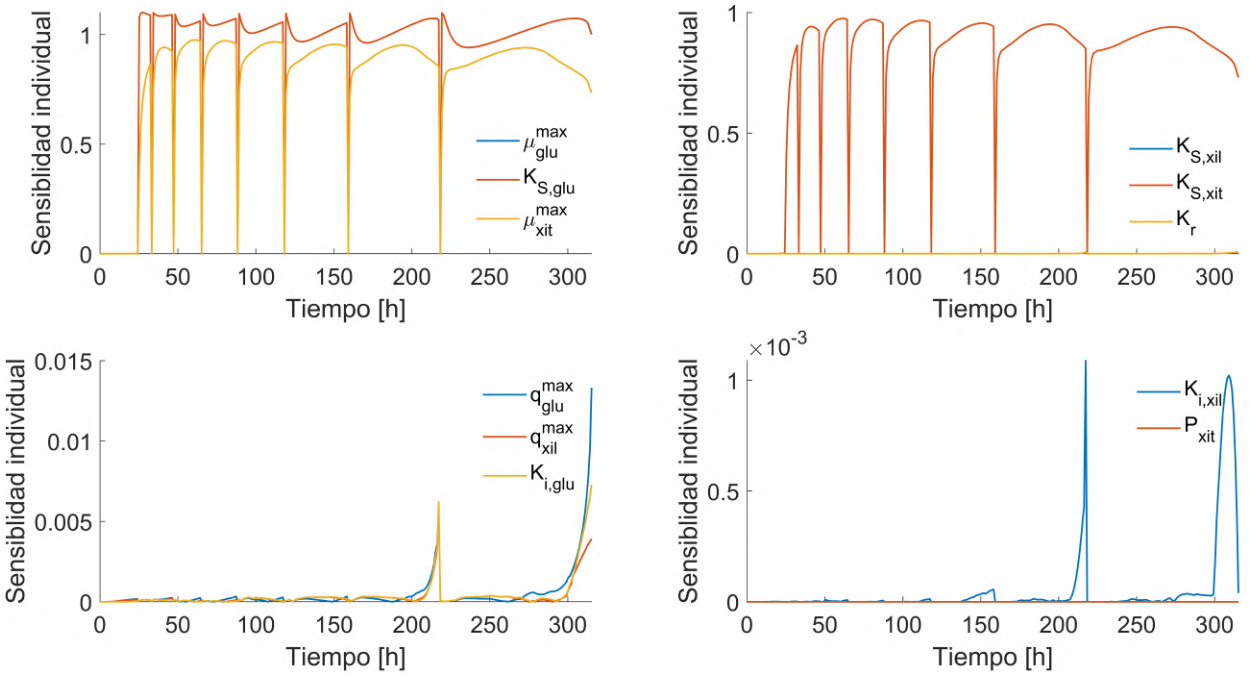


Figura C-12: Perfiles de índices de sensibilidad individuales de Sobol detallados con distribución normal para el modelo matemático de fermentación diaúxica para producción de xilitol.

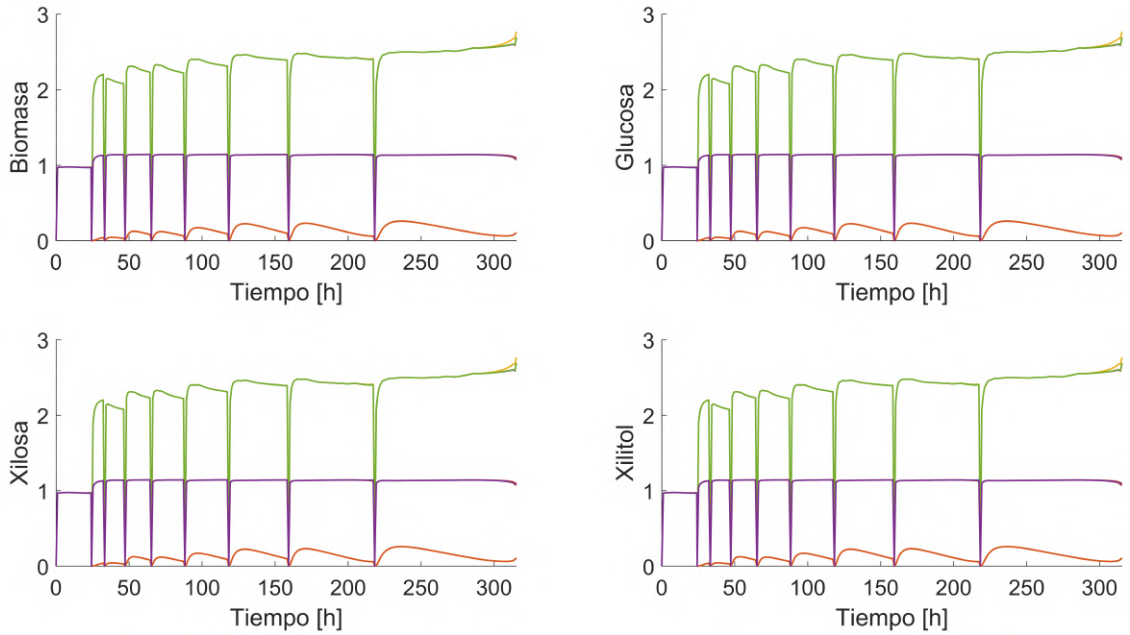


Figura C-13: Perfiles de índices totales de Sobol por estado para el modelo matemático de fermentación diaúxica para producción de xilitol con distribución normal.

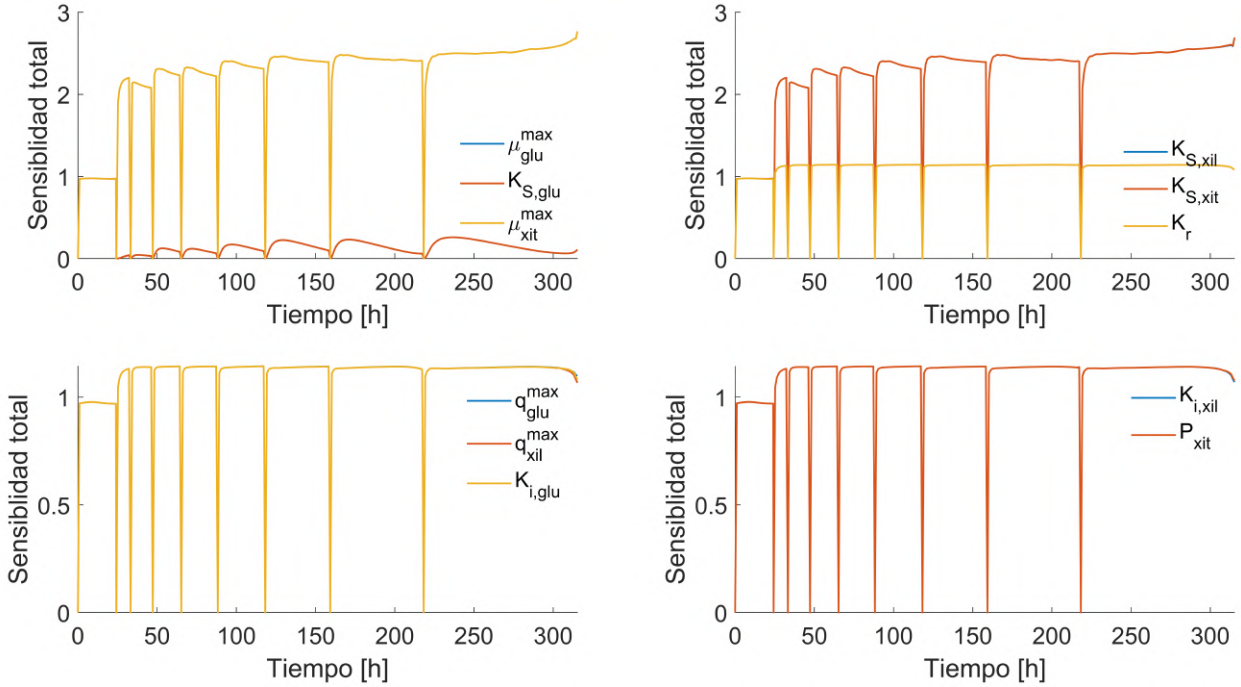


Figura C-14: Perfiles de índices de sensibilidad totales de Sobol detallados con distribución normal para el modelo matemático de fermentación diaúxica para producción de xilitol.

Bibliografía

Aarts, E. & Korst, J. (1989). Simulated annealing and Boltzmann machines. New York, NY; John Wiley and Sons Inc.

Aster, R. C., Borchers, B., & Thurber, C. H. (2005). Parameter estimation and inverse problems. Elsevier.

Chapra, S. C. et al. (2012). Applied numerical methods with MATLAB for engineers and scientists. New York: McGraw-Hill,.

Dallas, S., Machairas, K., & Papadopoulos, E. (2017). A comparison of ode solvers for dynamical systems with impacts. Journal of Computational and Nonlinear Dynamics, Special Issue on Dynamics of Systems with Impacts.

Englezos, P. & Kalogerakis, N. (2000). Applied parameter estimation for chemical engineers. CRC Press.

Kennedy, J. & Eberhart, R. (1995). Particle swarm optimization (pso). En: Proc. IEEE International Conference on Neural Networks, Perth, Australia (pp. 1942–1948).

Lillacci, G. & Khammash, M. (2010). Parameter estimation and model selection in computational biology. PLoS computational biology, 6(3), e1000696.

Pelletier, G. J., Chapra, S. C., & Tao, H. (2006). Qual2kw: A framework for modeling water quality in streams and rivers using a genetic algorithm for calibration. Environmental Modelling & Software, 21(3), 419–425.

Pitt, J. A. & Banga, J. R. (2019). Parameter estimation in models of biological oscillators: an automated regularised estimation approach. BMC bioinformatics, 20(1), 82.

Plischke, E., Borgonovo, E., & Smith, C. L. (2013). Global sensitivity measures from given data. European Journal of Operational Research, 226(3), 536–550.

Rangaiah, G. P. (2010). Stochastic global optimization: techniques and applications in chemical engineering. World Scientific.

Reynolds, C. W. (1987). Flocks, herds and schools: A distributed behavioral model. Computer Graphics, 21(4), 25–34.

Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. Computer physics communications, 145(2), 280–297.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & Tarantola, S. (2010). Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. Computer physics communications, 181(2), 259–270.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008). Global sensitivity analysis: the primer. John Wiley & Sons.

Sheikholeslami, R. & Razavi, S. (2017). Progressive latin hypercube sampling: An efficient approach for robust sampling-based analysis of environmental models. Environmental modelling & software, 93, 109–126.

Sobol, I. (1993). Sensitivity estimates for nonlinear mathematical models. Mathematical Modelling and Computational Experiments, 1(4), 407–414.

Spall, J. C. (2005). Introduction to stochastic search and optimization: estimation, simulation, and control, volumen 65. John Wiley & Sons.

Teschl, G. (2012). Ordinary differential equations and dynamical systems, volumen 140. American Mathematical Society.

Weber, F., Theers, S., Surmann, D., Ligges, U., & Weihs, C. (2018). Sensitivity analysis of ordinary differential equation models: Methods by Morris and Sobol' and Application in R.

Ye, M. & Hill, M. (2017). Global sensitivity analysis for uncertain parameters, models, and scenarios. En: Sensitivity analysis in earth observation modelling (pp. 177–210). Elsevier.