Classification Models for Progression of Chronic Kidney Disease within a Secondary Prevention Program

Sergio Páez Moncaleano
Statistics

Universidad Nacional de Colombia
Faculty of Sciences
Department of Statistics
Bogotá, D.C.
November 2019

Classification Models for Progression of Chronic Kidney Disease within a Secondary Prevention Program

Sergio Páez Moncaleano
Statistics

Final project Presented to obtain degree of
MSc Statistics

Advisor
Campo Elías Pardo
Ph.D.

Coadvisor
Mauricio Sanabria
M.D.

Universidad Nacional de Colombia
Faculty of Sciences
Department of Statistics
Bogotá, D.C.
November 2019

Title in English

Classification Models for Progression of Chronic Kidney Disease within a Secondary Prevention Program

Título en español

Modelos de Clasificación para la Progresión de la Enfermedad Renal Crónica dentro de un Programa de Prevención Secundaria.

Abstract: Loss of renal function has severe repercussions in patients' health and life quality. Using scientific tools to improve the knowledge of the disease and to prevent its progression on each patient could prevent terminal stages and even save lives. For a set of patients enrolled in a secondary prevention program, which aims to avoid reaching advanced stages of chronic kidney disease, we developed a complete statistical strategy: first, we described and prepared the data set. Then, we made groups of patients and afterwards we fit some classification models to understand such partition. Finally, we developed and estimation of the patients' future trajectory. We found that the classification models had good performance, with even 90% of good classification, also, that the estimation on the future trajectory seemed to be reliable, even in patients in which the model was not trained. Finally, an interactive tool was created in order to allow a real use of the results of this work in the diary medical care.

Resumen: La pérdida de la función renal tiene repercusiones significativas en la salud y en la vida de los pacientes. Con el uso de herramientas estadísticas es posible mejorar el conocimiento de la enfermedad y predecir el comportamiento de esta en cada paciente, haciendo viable prevenir etapas terminales e incluso permitiendo salvar vidas. En este trabajo se combinan técnicas estadísticas con conocimiento médico en nefrología para obtener una herramienta que ayude a los médicos a tratar y a tomar decisiones sobre sus pacientes. Para este fin, se tomó un conjunto de pacientes que pertenecen a un programa de prevención secundaria que trata de evitar la llegada a fases avanzadas de la enfermedad renal crónica y, primero, se desarrolló una estrategia estadística en la que inicialmente se describió y preparó la base de datos. Después, se formaron grupos de pacientes y se ajustaron algunos modelos de clasificación para analizar las particiones. Finalmente, se realizó una estimación de la trayectoria futura de los pacientes. Encontramos un buen desempeño de los modelos de clasificación, con hasta el 90% de buena clasificación, además, la estimación de la trayectoria futura dio resultados confiables, incluso en pacientes en los que el modelo no se había entrenado. Finalmente, se creó una herramienta interactiva para permitir el uso real de los resultados de este trabajo en la práctica clínica diaria.

Acceptation Note

Thesis Work
Approved

Jury
Luis Guillermo Diaz, PhD.

Jury
Nelcy Rodriguez, Mtr,

Advisor
Campo Elías Pardo, PhD.

Coadvisor
Mauricio Sanabria, M.D.

Bogotá D.C., November 30th 2019

Dedicado a

A Dios por darme la vida y por la oportunidad de culminar esta etapa con alegría. Por estar siempre tan pendiente de mí. Los títulos no valen nada comparado con lo que he recibido de ti.

A mi esposa, Ángela María, este trabajo es superfluo comparado con el valor que tienes en mi vida y todo lo que has dado por mí.

A mi papá, un ser único, con un discernimiento impresionante. Me ha enseñado cosas muchísimo más importantes que cualquier profesor y me ha mostrado con hechos lo que significa ser un hombre y un papá.

A mi mamá. Dar la vida por mí es poco, comparado con todo lo que ha hecho. Ella ha sido una luz en mi camino para encontrar lo que realmente es importante.

A mi familia por siempre estar ahí para mí. No más aguantarme ya es mucho decir.

Agradecimientos

Gracias al profesor Campo Elías Pardo por ser el tutor del trabajo y por ayudarme a lograr este cometido.

Un agradecimiento a RTS, especialmente, al Dr. Mauricio Sanabria por su ayuda en la consecución de las bases de datos. Gracias por enriquecer de tantas formas el trabajo realizado, por el apoyo brindado y por los nuevos horizontes que se abrieron para mí. Gracias a Santiago Velasco, quien con sus conocimientos estadísticos me orientó y apoyó en este proceso.

Agradezco, finalmente, a Jose, Gaby y Maya por su apoyo en los lenguajes de programación. A Ticas, Joe y Mariana U. por sus consejos gracias a sus amplios conocimientos en estadística. También agradezco a Clara Inés y a Rosita por su apoyo en el uso de LaTeX.

Contents

List of Tables

List of Figures

Introduction

The kidney is a fundamental organ for cleaning the body; it removes waste and fluid excess through urine. It also produces hormones that affect the function of other organs. As an example, a hormone produced by the kidney, called renin, helps to regulate blood pressure, and another hormone, called erythropoietin, helps to control the production of red blood cells. The body fluids are also balanced thanks to the process of excretion and reabsortion (National Kidney Foundation 2017).

Chronic Kidney Disease (CKD) is the progressive loss of renal function. Today, CKD is considered a public health problem, not only because the increased incidence and prevalence of the disease itself, but also because CKD is associated to additional serious diseases such as hypertension or diabetes, among others. In addition, despite the advances in quality care, the terminal state of CKD, called Renal Replacement Therapy (RRT), has a high mortality rate for patients (Bradbury et al. 2007, Suri et al. 2013).

The high prevalence of CDK, and its strong relation with cardiovascular disease, is a heavy burden for health systems around the world. That is the reason why international guides recommend early diagnosis of CKD as an effective approach to this problem (Atkins 2005, Codreanu et al. 2006, Levey et al. 2005).

Secondary prevention programs implement strategies for patients with initial CKD, that is, to detect and treat patients before RRT is required. In fact, early intervention has shown a positive effect in life quality and can also delay or even prevent RRT (Hu et al. 2012).

The test to measure the level of kidney function and to determine the stage of kidney disease is called the Glomerular Filtration Rate (GFR) and it can be calculated based on the results of the blood sample. Table 1 contains the relation between GFR and the severity of Kidney Disease. A GFR above 90 indicates a normal kidney function, below 60 means the presence of CKD, and, usually, if the GFR is below 15, there is a kidney failure. The latter means that the kidney does not have the capacity to clean correctly the blood and it is necessary to perform a RRT (National Kidney Foundation 2017).

TABLE 1. Stages of CKD and estimate percentage of kidney function based on estimated GFR.

Stages of Chronic Kidney Disease	GFR*	% of Kidney Function
Stage 1: Kidney damage with normal kidney function **(E1)**	90 or higher	90 -100 %
Stage 2: Kidney damage with mild loss of kidney function **(E2)**	89 to 60	89 - 60%
Stage 3a: Between mild and moderate loss of kidney function **(E3a)**	59 to 45	59 - 45%
Stage 3b: Between moderate to severe loss of kidney function **(E3b)**	44 to 30	44 - 30%
Stage 4: Severe loss of kidney function **(E4)**	29 to 15	29 - 15%
Stage 5: Kidney failure **(E5)**	Less than 15	Less than 15%

*GFR is the main variable to determine the level of kidney function. As kidney disease gets worse, GFR decreases. In parenthesis the abbreviations of each stage. Taken from National Kidney Foundation (2017)

However, there are clinical standards for measuring GFR. In figure 1, we can see implicit challenges for analyzing CKD because each individuals' shape of trajectory has different movements that show the large internal differences in GFR measure. Also, every individual has a pattern that is unique in comparison to other patients, which shows the external differences among individuals. In fact, by comparing individuals with similar basal GFR, some of them rapidly decreased in GFR until RRT is required, some decreased slowly and some were stable during the whole study. Also, there are missing values, represented as lines with jumps. This is situation that is very common in health studies.

Previous challenges imply a higher demand for the models. The first two challenges could be addressed by segmenting the population, which means to make groups of GFR trajectories that behave similarly, reducing variability inside each group and decreasing the complexity of models. On the other hand, the third challenge is transversal and has incidence in both clustering and model fitting, since such methods usually require complete observations. The main solution will be to select carefully the patients (add exclusion criteria), however, imputation will be also used as a solution.

From a statistical perspective, this problem could be addressed in order to improve the knowledge about CKD behavior by using several clinical variables to explain the progress of the disease. As GFR is the main variable to measure CKD, the idea is to predict some progression patterns that can help physicians to take more accurate preventive actions.

As it will be seen in the following sections, GFR, which is the core of this study, is a very complex measure in terms of its variability between patients. This makes more difficult to describe it adequately and to implement political decisions to benefit patients with CKD.

The first proposal of this work is to help the attendant physicians to take decisions over their patients' treatment, giving them a tool that allows them to have an idea of the current GFR progression and a possible future behaviour, based on the historical measures.

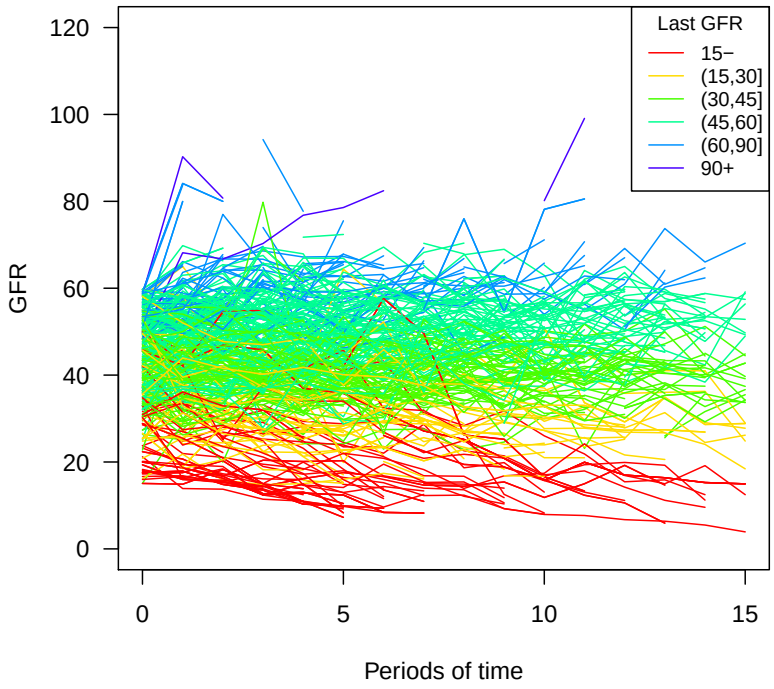


FIGURE 1. GFR trajectories sample during the time of the study. Colors indicate the stage of the last measure for each individual, according to table 1

The second objective is to generate knowledge for governments and health entities that helps them distribute resources appropriately, in order to invest on preventing CKD and its progression.

The third objective is to create a methodology that adequately describes CKD patients in terms of the progression of their GFR over time. This means to answer to previous proposals from a modern statistical approach.

The main idea is, first, to make groups of patients based on their GFR trajectories. Then, to use such GFR trajectories groups so that when a new patient arrives, the information of the entire dataset could be used to assign the individual to a specific group. The characteristics of each group will help to recognize aspects that could affect CKD progression; that way, the future trajectory could be computed.

This work is ordered as follows: the first chapter is dedicated to Statistical Methods, both to make groups of individuals and, given certain variables, to know to which group should a patient be associated. The second chapter combines the methods presented in the previous section in order to show the explicit methodology to analyse CKD. Finally, conclusions and future work are presented.

CHAPTER 1

Methods

1.1 General Framework

Assuming a set of N data points (individuals) $\{x_i\}_{i=1}^{N}$, each one has p features, it means that we have an $N \times p$ matrix X. Also, assuming that Y is the $N \times T$ matrix of outcome measurements, in this case, T represents the number of repeated mesures of GFR.

The covariance matrix of the matrix X is defined as

$$\Sigma_X = \frac{1}{N} \left(X - \mathbb{1}_N \mathbb{1}_N^t X/N\right)^t \left(X - \mathbb{1}_N \mathbb{1}_N^t X/N\right)$$

Where $\mathbb{1}_N = [1, 1, \ldots, 1]^t$ is the vector of ones of size N. Also, as $\mathbb{1}_N^t X/N = \bar{\mathbf{x}}$.

Which gives

$$\Sigma_X = \frac{1}{N} \left(X - \mathbb{1}_N \bar{\mathbf{x}}\right)^t \left(X - \mathbb{1}_N \bar{\mathbf{x}}\right) \tag{1.1}$$

Here the variance and the structure of the covariances or between the p variables are of interest. As in real life, the structure does not tend to be simple. An alternative approach is to look for smaller (much smaller than p) created variables that preserve most of the information given by this matrix Σ_X.

This is a well known problem in statistics and could be solved using principal component analysis (PCA). PCA basically uses linear combinations of x that retain maximum variance. The first linear combination would be to derive $\alpha_1^t x$ with maximum variance $var(\alpha_1^t x) = \alpha_1^t \Sigma_X \alpha_1$ subject to $\alpha_1^t \alpha_1 = 1$. With $\alpha_1 = (\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1p})^t$ a vector of constants. The solution of the optimization is to take α_1 as the eigenvector related to the highest eigenvalue. A complete review of PCA can be found in Jolliffe (2002).

Now, assuming that there are a total of K groups founded through the behaviour of Y, let G be the $N \times K$ matrix of class indicator variables ($\{G\}_{ij} = 1$ if and only if individual i is assigned to class j). Let also M be the $K \times p$ matrix of class means of X, that is, to compute the means of the p variables for each group separately. Finally let $\bar{\mathrm{x}}$ be the vector of means for the p variables of the whole set of individuals. Using G we could divide total variation as follows

$$\Sigma_X = \frac{1}{N}\left[(N-K)W_X + (K-1)B_X\right] \tag{1.2}$$

With

$$W_X = \frac{(X - GM)^t (X - GM)}{N - K} \tag{1.3}$$

and

$$B_X = \frac{(GM - \mathbb{1}_N \bar{\mathbf{x}})^t (GM - \mathbb{1}_N \bar{\mathbf{x}})}{K - 1} \tag{1.4}$$

W_X is the covariance matrix within groups and B_X the covariance matrix between groups. These matrices are the base of some of the following methods, guiding the path for partitioning data and developing the models.

Depending on the structure of data, we arrive to different but philosophically similar methods. If the goal is to use the input variables to predict the values of the output variables we are talking about supervised learning (Hastie et al. 2009).

Such task could be separated in two general groups depending on the type of response variable that wants to be predicted. The separation leads to name differently each task: We call *regression* when the output variable is quantitative and *classification* when we want to predict results of qualitative responses (Lebart et al. 1995, Hastie et al. 2009). In this work we will pay attention only to *classification* problems.

In the case in when the variable is observed longitudinally, meaning a variable which is repeatedly measured over time for each unit of analysis (sec. 1.3.3), the set of observations for each individual is called *trajectory* or *longitudinal observation* and the notation has an additional index: For the N trajectories, y_{it} is the value for the individual i on the time t, with $i = 1, 2, \ldots, N$, and $t = 1, 2, \ldots, T$, and T the longest possible time in which the individuals could be measured. Also, $y_i = [y_{i1}, y_{i2}, \ldots, y_{iT}]$ represents the entire trajectory of the individual i.

Based upon the mentioned above, the general structure of analysis is shown in figure 1.1. Both the document and the analysis per se will follow such structure.

In the following sections we are going to present the set of tools that will help us both to create groups of GFR trajectories (y_i) and explain the relation between clusters and a set of independent variables. As imputation will be a stage for preparing the data

FIGURE 1.1. Diagram of general steps proposed to analyze CKD data

set, in section 1.2 some theoretical details are shown. In section 1.3, we discuss how the groups are created; then, in section 1.4, groups are treated as response variable and methods for *classification* are presented. Afterwards, in sections 1.4.5 and 1.5, respectively, a discussion about prediction and model assessment is made. Every step will make sense since it respects and takes into account medical expertise in CKD.

1.2 Imputation

Longitudinal studies commonly deal with missing data. Also, as a clinical problem, studies deal with dynamic cohorts, which means that individuals could arrive late to the study and, therefore, have less measures than the rest of the cohort. Additionally, patients are in the right to abandon the study at any point for reasons that could vary from medical, to changes in their medical services.

As algorithms for clustering require completeness of the related matrix, in this case trajectories of GFR, we need to make an imputation of the related transition matrix.

1.2.1 Linear Interpolation

The linear interpolation replaces a missing value of the individual i at time t y_{it} by drawing a line between the two non-missing values that precede and follow the missing one. Let y_{ia} and y_{ib} be the closest preceding and following non-missing values of y_{it}, then

$$y_{it}^{LI} = y_{ia} - (t-a)\frac{y_{ib} - y_{ia}}{b-a} \tag{1.5}$$

With a and b the periods of the closest preceding and following non-missing values respectively.

1.2.2 Copy Mean

This method was created in Genolini, Écochard & Jacqmin-Gadda (2013) and implemented in **kml** package of `R` software (Genolini et al. 2015). It combines the linear interpolation (previously presented) and the population mean to adjust the imputation.

Let $\bar{y}$ be the mean trajectory of the population (in this case the mean GFR trajectory) as follows:

$$\bar{y} = \left(\frac{1}{N} \sum_{i=1}^{N} y_{i1}, \frac{1}{N} \sum_{i=1}^{N} y_{i2}, \ldots, \frac{1}{N} \sum_{i=1}^{N} y_{iT}, \right)$$

Let y_{it} be the missing value of the individual i at time t, let y_{ia} and y_{ib} be the closest preceding and following non-missing values of y_{it}, being y_{it}^{LI} the value of linear interpolation for time t to the mean trajectory $\bar{y}$. Then, the average variation (AV) is the difference between value of mean trajectory at time t, $\bar{y}_{.t}$ and the associated linear interpolation y_{it}^{LI} that is $AV = \bar{y}_{.t} - y_{it}^{LI}$

Copy mean imputes y_{it} by adding the AV to the linear interpolation, that is

$$y_{it}^{CM} = y_{it}^{LI} + AV \tag{1.6}$$

y_{iy}^{CM} the imputation by copy mean method is basically a correction of linear interpolation.

As it is possible to have missing values at the extremes of vectors (commonly called monotone missing values), the line joining first and last non-missing value of the respective individual is computed. Missing-value is replaced by fitted line evaluated at the first or last value of time.

1.3 Clustering Algorithm

The cluster analysis is useful for dividing the population, making decisions easier and more intuitively. In addition, this method is also used to generate descriptive statistics and to identify if data can be represented in a set of groups in such way that the degree of difference between the objects assigned to each cluster can be evaluated. This would be a descriptive rather than an inferential tool that helps to identify if it makes sense or not to perform statistical groupings (Hastie et al. 2009).

What encompasses the objectives behind cluster analysis is the notion of similarity (dissimilarity) between the objects that are going to be grouped. A cluster method would be formulated by the measure that allows it to group the objects that make up the base.

There are two types of clustering methods:

1. Those that allow to obtain a partition fixing the number of groups. Among the best known and widely the used is the K-means (Celeux & Govaert 1992).

2. Those that build a set of nested partitions, represented graphically as a tree or dendrogram, known as hierarchical classification.

Both types of methods have advantages and disadvantages, both require measures of similarity (dissimilarity) and distance between individuals and, as this document follows, both could be combined to obtain a better partition, even, they could also be combined with methods on principal axis (Lebart, Morineau & Piron 1995). General information about cluster analysis can be found in Everitt, Landau, Leese & Stahl (2011).

Before presenting the grouping algorithm, an introduction of the measures of distance or similarity will be made.

We also want to emphasize that, in this project, we created the clusters by using the trajectories of GFR, but not necessarily the trajectories of all individuals. This because we wanted to create groups using the most clean and reliable information as possible.

1.3.1 Measure of Dissimilarity

Assuming we have a matrix of observations Y of size $N \times T$ with N: The number of observations; T: The number of GFR measures, each object y_i is a vector of T components with $i = 1, 2, \ldots, N$. Thus, the distance between two observations y_i and y_i^a is given by

$$d(y_i, y_i^a) = \left(\sum_{j=1}^{p} (y_{ij} - y_{ij}^a)^m \right)^{1/m}$$

, called the Mikownski distance between the vectors. If $m = 1$, it is called the Manhattan distance and if $m = 2$, the Euclidean distance. Thus, a dissimilarity matrix D of size $N \times N$ is formed, which contains all the distances between the pairs of observations.

In practice, we use the Euclidean distance because of the good performance over real data and simplicity (Hastie et al. 2009), also, using the Euclidean distances for the following clustering methods, establishes inertia as a measure of homogeneity between groups and within groups, as presented in the next section. Then, clustering algorithms could be combined as both minimize the inertia within groups, given clusters with homogeneous individuals.

1.3.2 Inertia

The French School calls a measure of variability inertia (Lebart et al. 1995). It is a dispersion measure that could be used to understand, given a partition, how the set of classes varies internally and between in between classes.

Mathematically the inertia of the cloud of N individuals is:

$$Inertia_N = \sum_{i=1}^{N} w_i d^2(y_i, \mathbf{C}) \tag{1.7}$$

with w_i the weight of the individual i, such that $\sum_{i=1}^{N} w_i = 1$. Also, $\mathbf{C} = \sum_{i=1}^{N} w_i y_i$ the associated centre of gravity, usually weights $w_i = 1/N$ for all individuals. Which reduces the formula to

$$Inertia_N = \frac{1}{N} \sum_{i=1}^{N} d^2(y_i, \bar{\mathbf{y}}) \tag{1.8}$$

This equation coincides with the trace of the covariance matrix in equation 1.1, which helps to show that inertia is effectively a sum of variances, then, a measure of variability.

If we divide the population in K groups or classes, we could decompose the total inertia (variability) of equation 1.7 in two quantities: the variability inside of each group (intra classes inertia) and the variability between groups (between classes inertia). For a specific group $k : k = 1, 2, \ldots, K$ the intra class inertia would be:

$$Inertia_{g_k} = \sum_{i \in g_k} w_i d^2(y_i, \mathbf{C_k}) \tag{1.9}$$

$\mathbf{C_k} = 1/\mathbf{w_k} \sum_{i \in g_k} w_i y_i$ is the associated centre of gravity for group k, $\mathbf{w_k} = \sum i \in g_k w_i$ is the weight of class k, and g_k is the set of individuals in group k. When weights $w_i = 1/N$ for all individuals 1.9 to is reduced to:

$$Inertia_{g_k} = \frac{1}{N} \sum_{i \in g_k} d^2(y_i, \bar{\mathbf{y}}_\mathbf{k}) \tag{1.10}$$

$\bar{\mathbf{y}}_\mathbf{k}$ is the mean vector for individuals in group k. Then decomposition of inertia (variability) of the set of N individuals is:

$$\begin{aligned} Inertia_N &= \sum_{k=1}^{K} \mathbf{w_k} d^2(\mathbf{C_k}, \mathbf{C}) + \sum_{k=1}^{K} Inertia_{g_k} \\ &= \sum_{k=1}^{K} \mathbf{w_k} d^2(\mathbf{C_k}, \mathbf{C}) + \sum_{k=1}^{K} \sum_{i \in g_k} w_i d^2(y_i, \mathbf{C_k}) \end{aligned} \tag{1.11}$$

The first term of the equation (1.11) is the inertia between classes, and the second term is the inertia intra-classes. As it was said before, $d(\cdot, \cdot)$ is the Euclidean distance, then, the measure of dissimilarity between individuals is already selected. When $w_i = 1/N$ for all individuals, the centre of class $k = 1, 2, \ldots, K$ is $\mathbf{C_k} = \bar{\mathbf{y}}_\mathbf{k} = 1/\mathbf{w_k} \sum_{i \in g_k} w_i y_i$, and the

centre of the cloud of N individuals $\mathbf{C} = \bar{\mathbf{y}}$. Here we also found that inertia between and within classes is a trace of matrices in equations 1.4 and 1.3, respectively.

Partition of inertia will be useful in the following sections as minimizing intra class inertia would be a criterion for finding the partition for clustering procedures.

1.3.3 Longitudinal K-means

The K-means algorithm is one of the most prominent methods for grouping. It is a hill climbing iterative method belonging to expectation maximization (EM) class (Celeux & Govaert 1992). It starts with a set of initial points and in the following steps, points are reorganized and reassigned until stabilization. To asses the group membership it is common to use Euclidean distance (Hastie et al. 2009).

The method of K-means introduced by MacQueen (1967) was demonstrated to be an algorithm that decreases or minimizes the intra-class inertia (Lebart, Morineau & Piron 1995). Hastie, Tibshirani & Friedman (2009) presented the K-means algorithm as a hill climbing algorithm that solves an optimization problem of the total cluster variance. However it is different, in terms of equations that match and conclusions.

For K-means, it is necessary to set the number of groups (K) and the initial structure (starting values). Initial conditions have a crucial role over the performance on the algorithm; they affect the quality of the partition as it reaches local minimum.

Using the notation in Lebart et al. (1995), the K-means algorithm is structured as follows (in the notation superscript indicates the stage in the algorithm and the subscript the class):

Step 0 Set a K initial centres $C^0 : \left[C_1^0, C_2^0 \ldots, C_K^0\right]$. Compute the Euclidean distance between each individual and each one of the K centres $d(y_i, C_k^0)$. $i = 1, 2, \ldots, N$, $k = 1, 2, \ldots, K$. The individual i belongs to class g_k^0 if the nearest centre to the point is C_k^0. That gives a partition $P^0 : \left[g_1^0, g_2^0 \ldots, g_K^0\right]$

Step 1 Compute the K new centres of the classes $C^1 : \left[C_1^1, C_2^1 \ldots, C_K^1\right]$ taking the gravity centres of the partition of the step 0 $P^0 : \left[g_1^0, g_2^0 \ldots, g_K^0\right]$. These new centres induce a new partition $P^1 : \left[g_1^1, g_2^1 \ldots, g_K^1\right]$.

Step m Compute the K new centres of the classes $C^m : [C_1^m, C_2^m \ldots, C_K^m]$ taking the gravity centres of the partition of the step $(m-1)$ $P^{(m-1)} : \left[g_1^{(m-1)}, g_2^{(m-1)} \ldots, g_K^{(m-1)}\right]$. These new centres induce a new partition $P^m : [g_1^m, g_2^m \ldots, g_K^m]$.

The algorithm stops if the difference between the inertia of two consecutive iterations is smaller than a threshold, or because the prefixed maximum number of iterations is reached.

Usually, the partition depends on the initial selection of centres. There are several ways to initialize K-means: using random points, using individuals with maximum distance, or combinations between them; a review could be seen in Genolini et al. (2015). An additional way to initialize K-means is to use clustering methods that built nested partition as initial points (Lebart et al. 1995). We conducted Lebart's approach due to its benefits for the quality of the partition.

1.3.4 Hierarchical Clustering

The results of K-means algorithm rely on the election of the number of clusters and an initial point. In contrast, hierarchical clustering methods do not require such specifications. Instead, it is necessary to establish a dissimilarity measure between groups.

As the name may indicate, Hierarchical Clustering produces hierarchical representations in which each level of the hierarchy is a result of combining the groups of the next lower level. In the lowest level, each group contains a unique observation. On the highest level, there is just one group that contains all individuals.

Hierarchical methods are robust, meaning that a method applied to the same data set produces the same results and does not require a prefixed number of classes.

Strategies of hierarchical clustering are divided in two main groups: agglomerative and divisive. Agglomerative methods start at the bottom and, on each step, they merge a selected pair of clusters into a single one. This has as a result a grouping at the next higher level with one less cluster. The pair selected for the merging consists on the two groups with the smallest dissimilarity between groups. On the other side, divisive methods start at the top and, on each step, they separate recursively one of the existing groups at the level to create two new clusters. The partition is selected in order to obtain two new groups with the largest dissimilarity between groups. With both ways agglomerative and divisive, there are $N - 1$ levels on the hierarchy.

As mentioned before, those methods require a dissimilarity measure or distance between individuals. There are some of these measures on literature depending on the type of the variable and real application. In this project we are going to select (as presented before) the Euclidean distance.

Between the variety of possibilities of methods and the measures, we are going to present and use the hierarchical clustering method of Ward, because as K-means do, it minimizes the inertia within groups. Also, Ward's method could be easily combined with K-means to improve the quality of the partition.

1.3.5 Ward's Method

In order to obtain groups, Ward (1963) proposed to make classes that optimize an objective function and Wishart (1969) suggested the variance within groups (a measure of

variability) as the objective function. The french school called it inertia, which is basically a sum of variances.

For groups that have minimum intra-class inertia, the selected measure of dissimilarity must be the Euclidean distance. Also, as K-means do, classes join for the mimimum increasing amount of intra-class inertia. This corresponds to the Ward's method (Ward 1963, Wishart 1969).

To present the method, we followed the procedure presented in Lebart et al. (1995). The Ward's distance between groups A and B is given by:

$$W(A,B) = \frac{w_A w_B}{w_A + w_B}\, d^2(C_A, C_B) \tag{1.12}$$

Here w_A and w_B are the weights for the groups A and B, respectively. C_A and C_B are the centres of gravity of each group. As it was said before, $d^2(\cdot,\cdot)$ is the canonical Euclidean distance for two observations. This value is the increase of inertia within groups for mixing groups A and B in a single one. In a particular case, for two individuals i and l the Ward's distance is:

$$W(i,j) = (w_i w_j)/(w_i + w_j) d^2(y_i, y_j) \tag{1.13}$$

Then, for a set of N individuals, we can create a tree using the Ward's method as follows:

Step 0 Compute the matrix of Ward's distances between individuals using equation 1.13

Step 1 Join the pair of groups (individuals in the first step), which has the smallest Ward's distance, using equation 1.12. For the following steps the pair is represented by its arithmetic mean and the last elements are in total $N - 1$.

Step 2 Compute the Ward's distance between all individuals and the new group (equation 1.12).

Step 3 Erase the rows and columns corresponding to the individuals or groups joined and add one row and one column to record the distances between the new group and the others.

Step 4 Repeat the process until you reach one class.

1.3.6 Advantages and Disadvantages of Each Clustering Method

In essence both K-means and Ward hierarchical clustering are useful to obtain a partition. Both look for clusters with the minimum possible intra-class inertia and are naturally matchable. As it can be seen in table 1.1, methods are complementary, the lack of robustness of K-means indicates that applying the algorithm to the same data with different

starting points yield different results in terms of partition. This means that K-means only reaches a local maximum. In Ward's method, as the partition is nested, an observation that measures a closeness to the rest of the clusters gives clues to think that such observation should be assigned to a different group, but could not be changed. K-means is a fast algorithm as it requires low computational cost. On the other hand, by its hierarchical structures Wards' method does not need either initial points or a prefixed number of classes. In fact, by drawing the associated dendrogram we have a tool for selecting the number of groups.

TABLE 1.1. Comparison of advantages and disadvantages between Ward hierarchical clustering and K-means.

Characteristic	Ward	K-means
Robust	✓	X
Local minimum	X	✓
Low computational cost	X	✓
Does not require initial points	✓	X
Does not require number of classes	✓	X

Check mark means having a positive characteristic and "X" the opposite.

1.3.7 Combination of K-means and Ward

As seen in the previous section, both K-means and Ward have different characteristics and advantages that could be complementary used to strengthen one another.

K-means requires both initial points and number of clusters, which can be found by cutting Ward's dendrogram. Centroid partition define the initial points and, inmediatly. However, Ward's algorithm is rigid, since it doesn't allow to modify the set of individuals of each cluster, as K-means do. K-means iterations, on the other hand, allows to improve partition quality, making the algorithm maximal. As Ward's algorithm has always the same partition results, the initial points for K-means will not vary, which makes the algorithm robust.

The classification strategy is summarized in the following steps:

1. Compute a hierarchical classification with Ward's method over individuals.
2. Decide the number of classes and cut the tree.
3. Compute K-means of consolidation starting from the gravity centres of the partition obtained by cutting the tree.
4. Characterize the classes.

This procedure is implemented in package FactoClass (Pardo & Del Campo 2007) from software R (R Core Team 2019).

1.4 Supervised Classification

Supervised classification belongs to the context of supervised learning when the task is to predict qualitative variables. In that sense, several methods have been created to face such task; most of them depend on linear algebra. Then, assuming that the relation between response variables and independent variables could be modeled sufficiently well using linear structures, in the following subsections we present some prominent methods for classification.

1.4.1 Linear Discriminant Analysis

Linear discriminant analysis (LDA), developed by Fisher (1936) is a method created for the specific task of classification. It looks for a linear combination $\alpha^t X$ of the variables that maximizes the quotient of its between variance (eq. 1.4) to its within variance (eq. 1.3).

Then the linear combination $\alpha^t X$ has variances $\alpha^t W_X \alpha$ and $\alpha^t B_X \alpha$, and total variance $\alpha^t \Sigma_X \alpha$, using Σ_X as in equation 1.2.

Then we minimize

$$\frac{\alpha^t B_X \alpha}{\alpha^t W_X \alpha}$$

Which is the same as

$$\frac{\alpha^t B_X \alpha}{\alpha^t \Sigma_X \alpha}$$

Which is equivalent to finding a minimum to $\alpha^t B_X \alpha$ subject to $\alpha^t \Sigma_X \alpha = 1$. As seen in (Lebart et al. 1995), α is the eigenvector associated to the maximum eigenvalue of $\Sigma_X^{-1} B_X$.

As for the principal components, we can take the linear components corresponding to largest eigenvalue. There will be at most $r = min(p, K-1)$. The eigenvalues are proportions of the between classes variance, which could be useful to understand how many linear combinations to use.

The linear combinations found by this process are called the linear discriminant or the canonical variates. It is important to mention that with Fisher's the threshold between groups is not explicit. Then, in order to predict the class of individuals, it is a common practice to classify by choosing the group whose mean is nearest in the space of canonical variables.

Reference and details of LDA could be seen in (Venables & Ripley 2002, Ripley & Hjort 1996, Lebart et al. 1995).

1.4.2 Assignation of Classes in LDA

Assignation (or encoding) in classification is the task of assigning an observation to a group using a set of auxiliary features. In other words, if we have a new observation with p features $X_0 = [x_{01}, x_{02}, \ldots, x_{0p}]$ and an unknown response variable (membership category), we are going to assign a category for the observation X_0.

One of the most common ways to assign an individual to a group is to compute the distance between its information and the centres of each group.

Assume we have K groups, each one with a centre C_k, $k = 1, 2, \ldots, K$. Then, the projected variables are

$$Z_j = X\nu_j \ , \quad j = 1, 2, \ldots, d \tag{1.14}$$

With d as the number of canonical variables in LDA. With Z the matrix of canonical variables of size $N \times d$. The centre of the k^{th}-group is given by:

$$C_k = \frac{1}{n_k} \sum_{i:y_i \in g_k} [Z_{i1}, Z_{i2}, \ldots, Z_{id}] \tag{1.15}$$

Then, compute the projection of the individual over the reduced space

$$Z_j^0 = x_0^T \nu_j$$

Then, $Z^0 \in \mathbb{R}^d$ is the projection vector of the new individual over the reduced space. Finally, assign the new individual to the group in which the distance is minimal.

$$\text{Group individual }_i = argmin\ _{1 \leq k \leq K} \ \left\{ d(Z_1^0, C_k) \right\}$$

whith $d(\cdot, \cdot)$ the canonical distance over the transformed space Z.

Linear Discriminant Function

In the special case of LDA in section 1.4.1, they have created a boundary selection based on the Normality assumption, the following equation is equivalent to pre-fix boundaries.

$$\text{Group individual }_i = argmax\ _{1 \leq k \leq K} \ \{\delta_k(X_0)\}$$

With

$$\delta_k(X_0) = X^t \Sigma_{X_0}^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + log \pi_k \tag{1.16}$$

Here Σ_X is the matrix of variances of X and μ_k, the vector of means for the group k.

That is, for an individual, having a set of features X_0 we predict its response variable (assign the individual to a group) where the discriminant function is maximized. See Hastie et al. (2009) for details.

1.4.3 Logistic Regression

The logistic regression (LR) model arises from the desire to model probabilities for the K classes via linear functions in the independent variable X. LR uses log-odds of conditional probabilities to establish how data should be grouped (see 1.4.1). LR belongs to the well known family of Generalized Linear Models (Chatfield et al. 2010), when comparing the log-odds between classes we write:

$$
\begin{aligned}
log\frac{P(G=g_1|X=x)}{P(G=g_K|X=x)} &= X^T\beta_1 \\
log\frac{P(G=g_2|X=x)}{P(G=g_K|X=x)} &= X^T\beta_2 \\
&\vdots \\
log\frac{P(G=g_{K-1}|X=x)}{P(G=g_K|X=x)} &= X^T\beta_{k-1}
\end{aligned}
$$

Note that the model depends on $K-1$ log-odds in order to respect the constraint that the probabilities sum one. In this case, the model uses the last class as denominator, but the denominator choice, could be arbitrarily selected as long as it is always the same. G represents the response variable and the set of $\beta = (\beta_1, \beta_2, \ldots, \beta_{k-1})$ preserves the relation between response variable and independent variables. However it is called "Logistic Regression" this is a classification model whose response variable is categorical and produces probabilities

In practice, each individual will belong to the group in which log-odds is the highest. When comparing two classes K (the reference class) and l (another one) for a specific individual i with predictors x_i

$$
\text{if } \frac{P(G=g_l|X=x_i)}{P(G=g_K|X=x_i)} > 1 \text{ then } log\frac{P(G=g_l|X=x_i)}{P(G=g_K|X=x_i)} > 0
$$

Which means that for individual i, the probability to belong to class l is higher than the reference class, if all the log-odds are negative. That means that the reference class should be selected for such individual.

1.4.4 Supervised Principal Component Analysis

Barshan, Ghodsi, Azimifar & Jahromi (2011) proposed a supervised version of the well known PCA (Jolliffe 2002). This is a method of supervised learning using both quantitative and qualitative response variables (Barshan et al. 2011). The advantage of this method is that, theoretically, it captures any relation between variables, linear, quadratic, etc.

The idea of supervised PCA (SPCA) is to reduce dimensionality extracting principal components of the data that have maximal dependence to the target variable. Understanding dependence as any kind of relation (linear, quadratic, sinusoidal), even a relation that does not necessarily has an associated function.

Then, we address the problem of finding the subspace U^tX such that the dependence between the subspace and the response variable Y is maximum. In conclusion, the idea is to maximize:

$$tr(U^tXHLHX^tU) \quad \text{Subject to } \mathrm{U^tU = I} \tag{1.17}$$

Where $L = Y^tY$ and H is a projection matrix $H = I - 1/n\mathbb{1}_N\mathbb{1}_N^t$. Which has a closed form and, as solution of U, the eigenvectors of the matrix $Q = XHLHX^t$.

When Y is equal to the identity, the optimization becomes the classical pricipal component analysis PCA (Barshan et al. 2011, Jolliffe 2002).

1.4.5 Assignation of Classes

As in LDA the rule created with SPCA model is used to relate the class with the predictor variables. Assuming K groups, each one with a centre C_k, $k = 1, 2, \ldots, K$ and remembering that dependence in SPCA is maximized by computing eigenvalue decomposition for a projection matrix (Eq. 1.4.4), then

$$Z_j = X\nu_j \ , \quad j = 1, 2, \ldots, d \tag{1.18}$$

d is the number of eigenvectors selected to reduce the matrix in 1.4.4 and Z a matrix of principal components of size $N \times d$, then the centre of the k^{th}-group is given

$$C_k = \frac{1}{n_k} \sum_{i:y_i \in g_k} [Z_{i1}, Z_{i2}, \ldots, Z_{id}] \tag{1.19}$$

Computing the projection of the individual over the reduced space

$$Z_j^0 = x_0^T\nu_j$$

Then, $Z^0 \in \mathbb{R}^d$ is the projection vector of the new individual over the reduced space. Finally, we assign the new individual to the group in which the distance is minimal.

$$\text{Group individual }_i = argmin\ _{1 \leq k \leq K}\ \left\{ d(Z_1^0, C_k) \right\}$$

$d(\cdot, \cdot)$ is the canonical distance over the transformed space Z.

1.5 Model Assessment and Selection

Classification models are mainly compared using the percentage of good classification. A first measure called *Apparent* percentage of good classification could be computed and compared with cross validation to give an idea of the quality of the model.

It is quite simple to compute; the steps are the following:

1. Fit the model with all N individuals.
2. Predict the class for all N individuals.
3. Compare the prediction and the real membership.

This measure overestimates the proportion of good classified individuals because it uses the same observations to fit and predict. In order to avoid this weakness, the percentage of good classification could be computed using *Cross-Validation of N-folds*. This method is widely used and the estimation of the percentage of good classification is more reliable.

Which means following this steps:

1. Ignore the information of the i^{th} individual.
2. Estimate the model with the rest $N - 1$ individuals (See section 1.4).
3. Assign the i^th individual to a group (See section 1.4.5).
4. Compute steps 1 to 3 for $i = 1, 2, \ldots, N$.

As the real group values are known, the percentage of good classification could be computed by comparing the prediction made by previous algorithm against real values. The largest the percentage of good classification, the better the classification method.

Cross validation has several advantages. First it does not depend on creating a sample, in consequence, the percentage of well classified individuals is always the same. In the case of sampling, the estimation depends on both samples and repetitions, however, it is computationally heavier. For small populations it gives a better idea of the real performance of the classification algorithm. Also, as it is computed for each individual separately, it gives an idea of the behavior for new data in prediction tasks.

1.5.1 Variable Selection

In supervised learning, the main way to select models is by the percentage of good classification. We used methods of this section to make a variable's preselection, that would be tested later with cross-validation.

The way to preselect variables is the same one that is used to perform a one-way multivariate analysis of variance (MANOVA). MANOVA is useful to compare the linear composite of means between the K groups as it tests the null hypothesis that the means, on a set of related dependent variables, do not vary across different groups (Kent et al. 2006).

In supervised classification, the response variable is categorical; in one-way MANOVA is the predictor variable. In both cases, the relation between the categorical variable and the predictors (response variables in one-way MANOVA) is analyzed. Though, the problem is inverted in terms of the categorical variable of interest and the predictors, in both cases their behavior and relations are being studied. As seen in (Todorov 2007), if there is an important variable for understanding or explaining the categorical of interest and we remove it, we expect it will change drastically the one-way MANOVA statistic and viceversa.

This way, having K groups, and a set of p predictor variables $x = (x_1, x_2, \ldots, x_p)$

We conduct the following hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k \text{ given that } \Sigma_1 = \Sigma_2 = \cdots = \Sigma_k \tag{1.20}$$

Against

$$H_1 : \mu_i \neq \mu_j \text{ for certain } i \neq j \tag{1.21}$$

This basically means that all groups have the same mean and that the model doesn't explain correctly the variability of the data.

The most common statistics for MANOVA are summaries based on eigenvalues of the sum of squares and products (SSP) matrices:

- $\Lambda_{Wilks} = \prod_{i=1}^{p} \frac{1}{1+\lambda_i}$
- $\Lambda_{Pillai} = \sum_{i=1}^{p} \frac{\lambda_i}{1+\lambda_i}$
- $\Lambda_{LH} = \sum_{i=1}^{p} \lambda_i$

The sum of squares and products matrices are the following:

$$\text{SSP}_{\text{Total}} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ki} - \bar{x})(x_{ki} - \bar{x})^t$$

$$\text{SSP}_{\text{Within}} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (x_{ki} - \bar{x}_k)(x_{ki} - \bar{x}_k)^t$$

$$\text{SSP}_{\text{Between}} = \text{SSP}_{\text{Total}} - \text{SSP}_{\text{Within}}$$

λ_i is the eigenvalue of the matrix $\Lambda = (\text{SSP}_{\text{Total}})^{-1}(\text{SSP}_{\text{Within}})$. A comparison of methods could be seen in Warne (2014). We decided to use the Wilks' Lambda statistic.

If the null hypothesis is rejected, there is a difference between at least one pair of groups, and the variables are useful to explain the variability of the model.

Previous equations coincide with the formulation made in equations 1.2 - 1.4, as section 1.3.2 where we presented the inertia for the response variable Y. Here we use the same formulation for the set of predictor variables X, separating the total variability in two parts, one related to variability inside groups, the other with the variability between groups.

We could use previous methods as a variable selection technique. It is a common step-wise technique using Λ statistic. The idea is to implement several models and select the one with the lowest Λ (Todorov 2007).

Step 0: Fit the model with all the possible variables

Step 1: Compute the Λ_{Wilks} for the model

Step m: Fit the model with one less variable

Step m+1: Compute the Λ_{Wilks} statistic for the reduced model.

Step m+2: Erase the variable if the Λ_{Wilks} increases or the p-value increases

The previous algorithm is the backward selection method, that will be used for the following steps. The forward and step-wise methods are basically the same with a different direction, details can be seen in Todorov (2007). Step-wise selection using Wilks' Lambda is implemented in klaR package (Weihs et al. 2005).

CHAPTER 2

Implementation

In this chapter we present the stages to implement the statistical methods presented in the previous one. In section 2.1 we present the data set, the set of variables of interest and some measures of them. Then, in section 2.2, we present the necessary transformations and decisions to obtain the data set for statistical procedures. Afterwards, in section 2.3, we perform the clustering of trajectories using the GFR trajectories and, finally, in section 2.4 we fit the supervised classification methods and, based on different measures, we select a final structure for the models to generate the results and the interactive tool for physicians.

2.1 Preliminary CKD Data Set

Following the steps presented in figure 1.1, this section contains the original data set and the stages to obtain the matrix to apply both the cluster algorithm and the supervised models.

This section is organized as follows: in subsection 2.1.1 information of original data set is found; subsection 2.1.2 specifies some data structures of the GFR variable. In subsection 2.1.3 information about the cause of patients' follow-up ending can be found, as well as subsection 2.2 contains some methodological decisions that were applied in the data set in order to accomplish the objectives and guarantee certain data structure. Subsection 2.1.4 contains information of the additional variables different from GFR and cause patients' follow-up ending, that will be useful for analysing data; finally, in subsection 2.2.4 the methodological decisions to obtain the data set to be used for the following sections are presented.

All these set of subsections will allow the reader to replicate the methodology and to carry out such a task. We strongly recommend reading all subsections in order.

2.1.1 Generalities

The data was provided by Renal Therapy Services Latin America (RTS) and the study has the structure of dynamic cohort, since patients could be admitted to the program at any moment. It is an observational study of historic cohort. The entire patients cohort was admitted between January 1^{St} of 2009 and December 31^{St} of 2014 (a duration of 2190 days or, equivalently, 73 months). Inclusion criteria were: at least one year of active permanence in the program, at least 18 years old; entrance to the *Health Renal Clinic Program* with Glomerular Filtration Rate (GFR) between 15 and 59 ml/Min in the first measure; and to have at least four measures of GFR measures throughout the study time. The only exclusion criterion, due to the clinical advice, was that individuals with advanced cancer could not be admitted. Previous description gives a total of $N = 3048$ patients in the program. All patients signed an informed consent document and all information about them was anonymized in the data set thanks to the use of codes, which made it impossible to identify any participant. This study was reviewed by an ethics committee to guarantee the safety of the patients.

2.1.2 Longitudinal GFR

The disease's progression of a certain patient is measured by the estimated GFR and the stages are presented in table 1 (Levey et al. 2005). An individuals' sample is presented in table 2.1. The structure for all the $N = 3048$ patients is the same, so presenting this subset is enough to understand the rest of them.

Each patient has a unique anonymous identification called *Code Key*, and the only one who can use this code to get sensitive patient data is RTS.

The dates on the second column indicate the day when the patient went to the medical centre, biological samples were collected, and the GFR was estimated. The first date when the patients were admitted to the program and the first GFR was estimated. Following dates indicate the follow-up of the disease. Depending on the medical decision dates are separated between 3, 4 or 6 months. The last date is the last measure of the patient. In the next subsection, an additional variable, called $Drop - Out$, will be explained. This variable gives information about the cause of the end of the follow-up. The third column is the corresponding stage using table 1. The fourth column is the estimation of the GFR.

Patient 2 was admitted on September 26^{th} of 2011, with previous GFR measures, and was measured 4 times every 3 months mostly, remaining between stage $E3b$ and $E3a$. Patient 3 was admitted on May 5^{th} of 2010 and was measured 8 times, mostly with a month's separation, but once with a separation of 6 months and another time with a separation of 4 months between medical appointments, remaining most of the time in stage $E3b$. Finally, patient 4 was admitted on August 12^{th} of 2012 and the last measure was in January 1^{st} of 2014; the pattern has 8 measures and, most of the time, the patient remained in stage $E3b$.

TABLE 2.1. Stage values and GFR for each consultation date for individuals 2, 3 and 4

Code Key	Consulting Date	Stage	GFR
2	26/9/2011	E3b	37.2
2	27/12/2011	E3a	45.1
2	21/3/2012	E3a	48.6
2	25/6/2012	E3a	45.3
3	5/5/2010	E3a	41.9
3	12/8/2010	E3b	38.9
3	17/11/2010	E3b	41.2
3	24/12/2010	E3b	35.4
3	22/2/2011	E3b	32.7
3	16/3/2011	E4	25.0
3	2/10/2011	E3b	33.5
3	23/1/2012	E4	24.4
4	12/8/2012	E3	40.9
4	30/10/2012	E3b	39.0
4	4/12/2012	E3b	39.0
4	15/1/2013	E3b	40.6
4	19/4/2013	E3b	41.4
4	15/7/2013	E3a	45.4
4	21/10/2013	E3b	40.0
4	28/1/2014	E3b	37.7

2.1.3 Drop-Out Variable

Using this variable we can identify the reason why a patient's disease follow-up ended. This will be helpful for both describing the data set and deciding which individuals will be selected for the following procedures.

Table 2.2 contains the causes why patients' disease follow-up ended in percentages. Most patients have as most common cause of follow-up ending the *End of Study*, which means that, though the study has ended 70.5% of patients still remain on it. As it is a dynamic cohort, not all individuals have measures on the $T = 15$ periods, for the time of a patient starts when entering the study. Additionally, undesirable clinical causes such as RRT, Death, Palliative Care, and Kidney Transplant sum 7.3% of the $N = 3048$ patients.

TABLE 2.2. Total and percentage of patients by causes of end of follow-up in the secondary prevention program.

Cause of end of follow-up	Percentage
End of Study	70.5
External Consultation	11.3
Loss of follow-up	4.7
RRT	3.7
Change of Insurer	3.4
Death	3.0
Abandonment of treatment	2.8
Palliative Care	0.5
Kidney Transplant	0.1
Total	100.0

2.1.4 Additional Variables

Table 2.3 contains the set of variables that were measured in the set of individuals; the majority are clinical variables that may be useful for supervised classification algorithms as predictor variables. Table 2.3 also contains summary statistics for demographic variables and for the most important clinical variables.

On the set, a total of $N = 3048$ individuals have CKD distributed as following: 47.4% are male, between 21 and 97 years, with an average age of 70.1 years. 52.5% are women, between 21 and 96 years, with an average age of 70.7 years. At the beginning, patients on data set have a CKD on stages lower than 5, that is, GFR less or equal to 60 ml/min and greater or equal to 15 ml/min.

During a review, we found that most variables could not be used for the following steps due to high levels of missing values. In the end, remaining variables are:

GFR, Age, Gender, Dx_DM, Dx_HTA, BMI, DiasBP and SysBP

Creatinine, Drop-Out and Cause could not be used as they are represented with other variables or are theoretically useless for modeling. Creatinine is used to compute GFR; Drop-Out is measured at the end of the study and we do not have an interest in forecasting if the individual will finish or not the study; and the Cause of CKD depends on variables like D_DM or D_HTA.

TABLE 2.3. Summary statistics and abbreviations for clinical and socio-demographic and clinical variables.

Baseline characteristics	Values		Var. Abbr.	% Missing
Female sex (n,%)	1285	52.5	Sex	0.0
Age (mean, SD)	70.4	11.5	Age	0.0
School level (n,%)			School	0.5
Illiteracy for reading and writing	165	6.7	*School1*	
Elementary	563	23.0	*School2*	
High School	1479	60.4	*School3*	
Technical, university or graduate	238	9.7	*School4*	
CKD cause (n,%)			Cause	1.60
Hypertension	1455	59.5	*CauseHTA*	
Diabetes	439	17.9	*CauseDM*	
Autoimmune	81	3.3	*CauseInm*	
Other	431	17.6	*CauseOth*	
Unknown	39	1.6	*CauseUnk*	
History of cardiovascular disease (n,%)	101	4.1	Cardi_Past	95.8
Diabetes diagnosis (n,%)	618	25.3	Dx_DM	0.0
Hypertension diagnosis (n,%)	2085	85.3	Dx_HTA	0.0
Glomerular filtration rate ml/min/1.73m^2 (mean, SD)*	46.8	9.7	GFR	10.2
Body mass index (mean, SD)	26.7	4.3	BMI	0.0
Systolic blood pressure mmHg (mean, SD)	133	20.1	SysBP	0.0
Diastolic blood pressure mmHg (mean, SD)	73.8	11.4	DiasBP	0.0
Albumin gr/dl (mean, SD)	4.3	0.4	Albu	73.8
Hemoglobin gr/dl (mean, SD)	14.1	1.8	Hemo	48.1
Blood Urea Nitrogen mg/dl [median;IQR]	26.6	12.1	BUN	44.0
Uric acid mg/dl (mean, SD)	6.7	1.7	UricAcid	76.2
Glycosylated hemoglobin % [median;IQR]**	7	1.9	HemoGly	90.9
Proteinuria gr/day [median;IQR]	0.1	0.2	Prot	98.5
Cholesterol LDL mg/dl [median;IQR]	112.4	46.8	LDL	82.0

$N = 2445$

* GFR is the main variable for CKD, as the observation is a trajectory of GFR measures. The missing value is computed by the quotient of the sum missing values inside each vector divided by the sum of the length of all vectors. ** Measured in diabetics only.

2.2 Data Set for Statistical Procedures

In order to achieve the proposed objectives and develop the analysis, it is necessary to first make a set of transformations to the original data, particularly to the GFR. Each stage follows both statistical and clinical expertise.

In figure 2.1, the changes and a brief explanation are shown. After the Cut-Off stage, we pass from $N = 3048$ to $N = 2445$ patients in order to reduce the amount of noise due to imputation, that is, to have as much as real information measured. Also, we decided to work with individuals that remained during the entire study, or droped-out due to CKD such as $Drop-Out$ of RRT, kidney transplant, palliative care or death. That gives a final total of patients of $N = 386$.

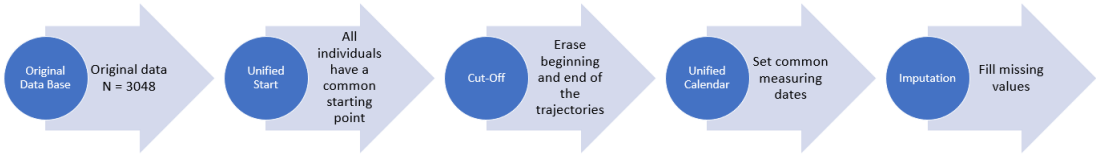


FIGURE 2.1. Diagram of changes to obtain the matrix for statistical procedures

2.2.1 Unified Start

The first methodological decision was to set aside the calendar. This way, all individuals will begin the study in a common initial time that was called day zero, which corresponds to the day they entered the study (first date of consultation). From this common starting point, the following measures were recorded to form the GFR trajectory of each patient. From this, the duration of each individual on the study and the variability of the pattern could be studied.

According to this, in practice there are three types of patients:

1. Patients who completed 2190 days, that is, they stayed during the entire study.
2. Patients who started at the beginning of the study but left prematurely, due to the causes described in table 2.2.
3. Patients who started after the beginning of the study.

In practice, these three types of patients describe the entire population. In terms of the right censoring, it is the same to leave the study prematurely than arriving late, however special attention must be paid to patients who left the study due to progression of CKD (like RRT or transplant). This decision implies that all censoring is charged to the right, which will be studied in the following subsection.

2.2.2 Cut-Off

Clinically, it is known that incident CKD patients have more erratic GFR patterns (Levey et al. 2005). Sometimes, because incorrect diagnosis is made and the patient doesn't have CKD in reality but an isolated kidney malfunction. Sometimes, because the clinical treatment has not regulated the kidney function and therefore the actual state of the patient's disease is unknown. Both phenomena cause instability in the GFR measures and noise in patients information, this justifies trimming the trajectories appropriately.

A histogram of stage changes through time is shown in figure 2.2. This is a histogram of absolute frequencies of the number of stage changes, according to table 1. It could be seen that the first 1200 days of the study collected approximately 27% of the total transitions presented throughout the study, while in the last 245 days the cumulative percentage of recorded transitions is barely close to 0.98%. As mentioned before, first part has high instability in CKD, also, the last one has high levels of right-censoring.

The previous facts lead to a second methodological decision: to cut the study time by removing the first 120 days and the last 240 days, reducing the study to 1820 days. This prevents considering patients that might not have CKD, but another kind of acute illness, and, additionally, since the last periods are high censoring, noisy information is avoided.

Cutting the study time also meant reducing the number of individuals observed from $N = 3048$ to $N = 2445$ as some of them did not meet inclusion criteria after the transformation. This cut was made with the purpose of fulfilling the inclusion criteria that dictated that, at the beginning of the study, the individuals should be in stages 3a, 3b, or 4. Thus, in the new zero point of the study time, those individuals who did not meet this criterion were removed. Particularly patient "2" in table 2.1 was removed.

2.2.3 Unified Calendar

After a unified start was set and the cut-off was applied, a third methodological change was applied. The motivation was the following: clinical expertise indicates that usually individuals with CKD are measured every 3 or 4 months because that period is enough to identify changes due to disease and not to randomness. Then, common periods of 120 days were created in order to have common days for repeated measures.

Ideally, patients should go to the physician on fixed dates, however, it is common to have delays or longer separations between measures. In that sense, in order to homogenize measure dates, an unified calendar was suggested, as table 2.4 shows.

It means that we have a matrix of $N = 2445$ patients with at most $T = 15$ measures (initial measure plus 15 periods of 120 days). All individuals have GFR at $t = 0$, however, later we show that different number of patients will be used for different methodological steps.

In summary, there were three changes

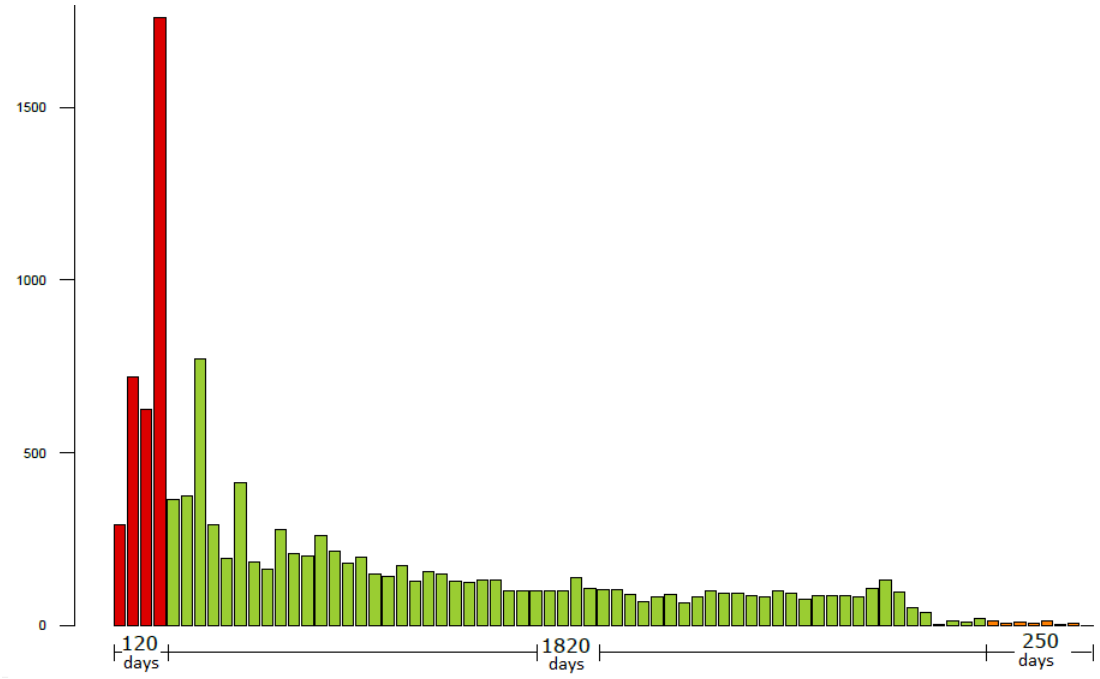


FIGURE 2.2. Histogram of absolute transition frequencies between stages of the CKD

TABLE 2.4. Algorithm for unified calendar

Step	Description
0	Measure at time $t = 0$ is the first GFR estimation. Following measure will be the nearest to the date after 120 days.
1	If there is no measure between the interval $(120 - 60, 120 + 60) = (60, 180)$ that period will have missing value.
⋮	⋮
	Measure at period t =1,2,..., 15 will be the nearest to the date after 120*t days.
t	If there is no measure between the interval (t*120-60 , t*120+60) that period will have missing value.

1. Assigning the first GFR measure as the starting point in order to have a common initial time, called time zero.

2. Cutting the tails of the follow-up period (starting 120 days and lasting 250).

3. Set common measure periods of 120 days.

The previous methodological decisions induce the table 2.5 which is equivalent to table 2.1 after the preprocessing stages. As it can be seen, individual "2" was eliminated because after preprocessing he didn't satisfy the inclusion criteria. Also both individuals "3" and "4" lost their first measure. In the table we have different cases: measures that are lost because periods do not capture such fine measures and measures that have missing values because any value was inside the period.

The table helps us to identify two facts after the methodological decisions: first, some measures are lost because of the creation of periods, however, from clinical expertise it is

not important as the GFR would be more realistic in terms of CKD behaviour. Second, these decisions help us to have an easy to describe (by periods) and to visualize matrix. For example, periods without measures in table 2.5 are not as easy to identify on its original form like in table 2.1.

TABLE 2.5. Values of GFR for each period of time for individuals 3 and 4.

Code Key	Period	Days	GFR
3	0	0	38.9
3	1	120	35.4
3	2	240	25.0
3	3	360	-
3	4	480	24.4
4	0	0	39.0
4	1	120	41.4
4	2	240	45.4
4	3	360	40.0

Symbol "-" means no measure for such period.

Although the previous decisions had clinical support, they induced an additional difficulty: statistical methods like linear discriminant analysis (LDA) or K-means can not deal with missing data. This means that the next step is to use imputation methods to fill such blanks.

2.2.4 Imputation

After implementing some preprocessing techniques to obtain a matrix with a strategic structure, the idea is to obtain a table that is useful for the following statistical methods.

As seen in table 2.1 and 2.5, the number of measures could drastically change among individuals due to several reasons like medical decisions or late arrival to the study, among others. For steps like clustering using Ward's method or discrimination with LDA (see Sec. 1) we needed to have a matrix without missing values. Then, in the cases in which individuals had missing values it was necessary to implement an imputation method (see Sec. 1.2).

It is clear that fewer imputed data is better. In that sense, the first criterion was to select individuals with measures from T_0 to T_{15}. This set of complete trajectories could only have missing values in the middle measures of the longitudinal pattern, as patients visit the physician yearly or every six months. Then, for this group we imputated values within the pattern and no future values.

The previous decision helps us to have the cleanest data possible in order to avoid noise for posterior methods. However, it should be noted that another part of the data is also as important as the complete one and is based on the Drop-Out variable described in section 2.1.3. Patients whose drop-out from the study is caused by the entrance to RRT have very important information for the progression of CKD. Without them, the method would be biased and incomplete. Then, the second decision was to induce patients whose

drop-out from the program could be affected by CKD. More specifically, patients who droped-out due to RRT, palliative care, death or kidney transplant.

For this second group of patients imputation methods would also complete the future pattern, however, it is not important as they usually have a systematic decrease.

In that sense, the target patients for the following methodologies had to participate in the entire duration of the program or abandon the program for undesirable clinical causes that impede the continuation of the follow-up. This gives us a total of $N = 386$ patients, who will participate in the next steps of Clustering (See Sec. 1.3), Supervised Classification (See Sec. 1.4), and methods in between.

In the following sections, we are going to present how the methods were implemented, decisions and challenges among them. As it will be seen, the steps, though they follow statistical decisions, are also guided according to clinical expertise. This allowed us to design a realistic and well grounded methodology.

It must be said that we implemented the imputation over all 2445 patients, as we are going to use such trajectories in specific stages, like evaluating the reliability of the model.

2.3 Clustering Trajectories

In this section we describe how to generate the groups of GFR longitudinal data using the methods of section 1.3. For such task we used the following packages: FactoClass for the combination of both K-means and Ward's hierarchical method using principal component analysis (PCA) as an additional tool (Pardo & Del Campo 2007, Jolliffe 1982) and kml (Genolini et al. 2015), which was mainly created for K-means but here it is used for more specific tasks such as imputation or graphical tools.

2.3.1 Principal Components Analysis Over the GFR Trajectories

Using all the $N = 386$ patients, the clustering method is applied. Following Pardo & Del Campo (2007), we compute a PCA over the matrix of $N = 386 \times T = 15$ periods. As all columns are GFR measures, variables have the same scale and are directly comparable, so a non-normed PCA was performed.

In the figure 2.3, on the left panel, the first principal component keeps 88.2% of the variability, meaning that it is a size factor, then, GFR measures can be summarized in a single component. However, we are going to use all components; (something equivalent to working with original variables). Experiments using a different number of axes were performed and we did not find relevant changes in the groups. On the right panel of the same figure we presented the variable projection over the first principal plane: the nearest arrows indicate a stronger correlation. That is why they are organized according to time periods. Also, the length of the arrows represents the approximate variance. In that sense,

periods have similar variance and tend to be shorter for the first periods (due to exclusion criteria).

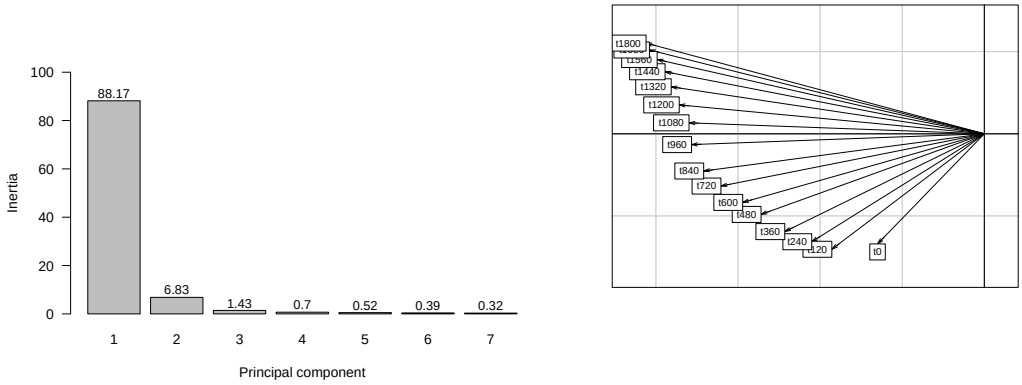


FIGURE 2.3. Left Panel: Inertia of the first seven principal components (equivalent to plot eigenvalues of the PCA). Right Panel: Plot of the first factorial plane for the variables of GFR measures.

2.3.2 Hierarchical Clustering

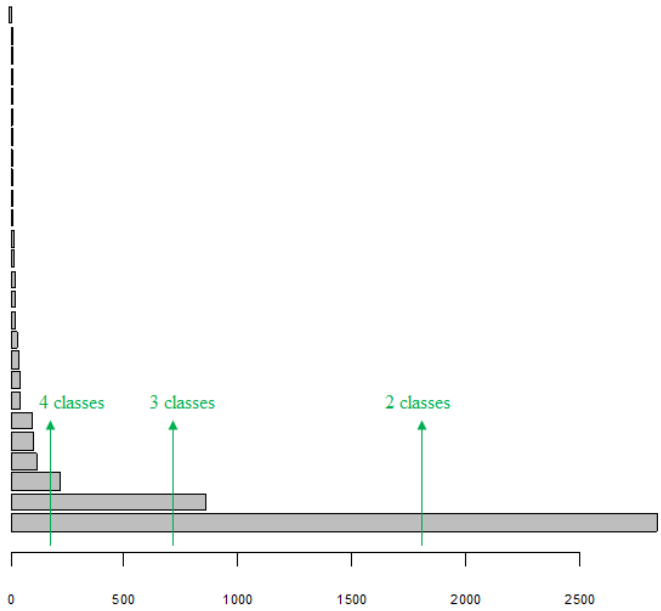
Ward's method uses such principal components for the hierarchical clustering. And K-means algorithm, on the other hand, uses Ward's partition as initial points to optimize such partition. Here we present the results after optimization, which is how we propose to divide the trajectories.

We plot in figure 2.4 the Ward's indexes and the table with the percentage of total variance for each number of partitions.

If we divide the set of $N = 386$ trajectories in two classes, the 58% of the inertia becomes between classes inertia. If we increase the number of partitions, then, the percentage of total variability explained by the between of classes inertia increases. Large jumps between bars indicate a large addition of inertia. Then, we propose to divide the $N = 386$ trajectories in four classes.

From both statistical and clinical perspectives, $K = 4$ groups are suitable to correctly divide the population. From medical perspective, the number of groups is important in the sense that physicians need to understand and adopt each group. Too many groups will be undesirable, for they affect the statistic parsimony. Also, understanding that as the quantity of groups increases, the complexity of prediction does too, but the size of each group decreases.

Additionally, when there are $K = 4$ groups, the percentage of variance explained by between inertia is around 80%, which is an important percentage.



K	% of Tot.Var.
2	58.2
3	75.9
4*	80.4
5	82.7
6	84.8
7	86.8
8	87.7
9	88.5
10	89.3

FIGURE 2.4. Left Panel: Histogram of Ward's indexes. Right Panel: Table of the percentage of total inertia explained by inertia between classes for 1 to 10.

2.3.3 Characterization of Classes

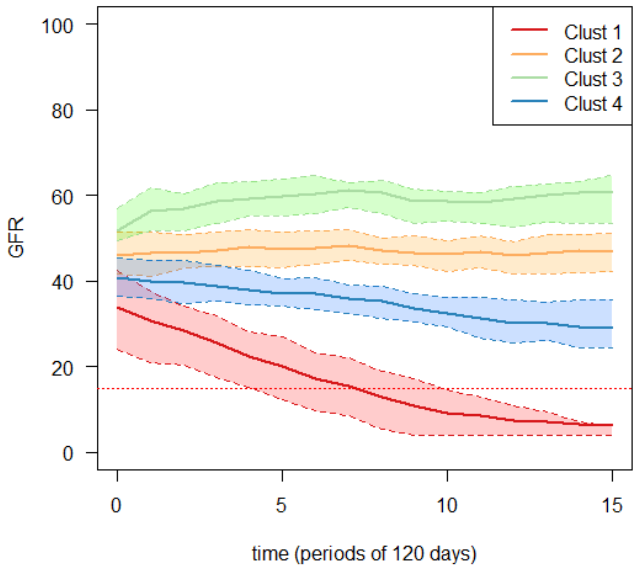
In figure 2.5 we presented the point-wise mean by cluster separately. In that sense, patients in *Cluster 1*, which rapidly decrease in GFR, would be a target for additional care and attention. On the other hand, individuals inside *Cluster 4* also tend to decrease in GFR, but not as fast as patients in *Cluster 1*. Additionally, patients in *Cluster 2* and *Cluster 3* tend to be stable, the main difference that remains is that the second one has higher GFR measures.

Additionally, in figure 2.6, we show the projection of $N = 386$ individuals over the first two principal components. We can see that a clear linear division between each one of the $K = 4$ groups exists: in the right, there are individuals with more advanced CKD in cluster 1, in the left, patients with the highest GFR trajectories. Both graphs indicate an order between groups in terms of the CKD progression measured by GFRs.

In tables 2.6 and 2.7 we present the characterization of the patients inside each cluster for qualitative and quantitative variables, respectively. It is important to note that such variables were not used to create groups.

Compared with the global measures, on each cluster we have:

- Cluster 1: High percentage of patients with diabetes, low percentage of patients with hypertension, and high percentage of patients with less than 65 years. For quantitative variables, the means of systolic and diastolic tension are higher than global means.



k	% n_k
1	24.9
2	28.8
3	20.6
4	25.7

FIGURE 2.5. Left Panel: Point-wise mean for each class. Horizontal dashed line is $GFR = 15$, which is a threshold that allows nephrologists to review the possibility of RRT.. Envelopes are quantiles 25 and 75 inside each group. Right Panel: Table of the relative size of each class.

- Cluster 2: Low percentage of patients with diabetes and high percentage of patients with hypertension compared to the global percentages. For quantitative variables, the means of systolic and diastolic tension are slightly lower than the global means.
- Cluster 3: Low percentage of patients with diabetes and low percentage of patients with hypertension. For quantitative variables, the mean of BMI is the lowest not only in comparison to the rest of the groups, but to the global mean.
- Cluster 4: High percentage of patients with diabetes (not as high as in cluster 1) and high percentage of patients with hypertension compared to the global percentages. For quantitative variables, the mean of each variable is basically the same as the global means.

2.4 Supervised Classification

In this section we present the results of fitting classification methods. The response variable is the membership class (previously computed by clustering) and the independent variables are the set presented in section 2.1.4: Age, Gender, Dx_DM, Dx_HTA BMI, DiasBP, and SysBP. Additionally, the first GFR measure and the difference between the first and the last measure, that is GFR_0, $(GFR_0 - GFR_{14})$ and $(GFR_0 - GFR_{15})$.

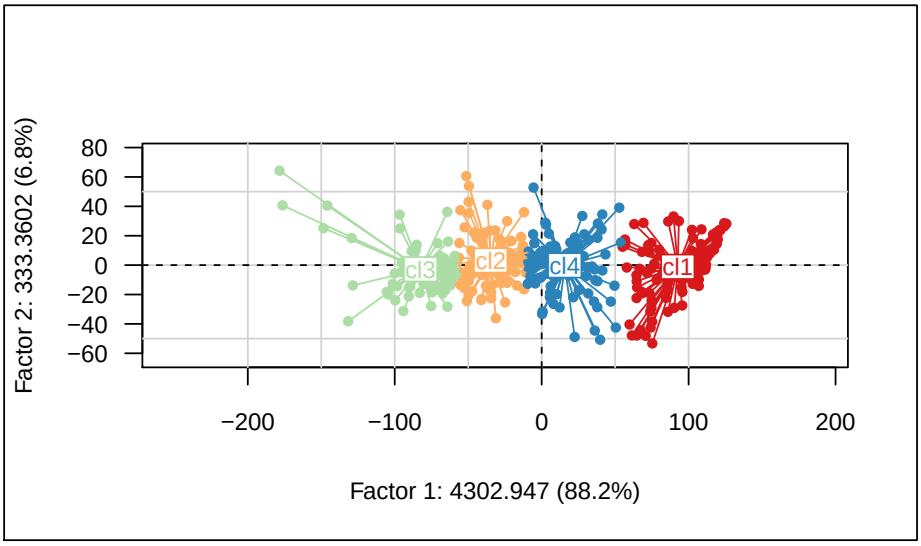


FIGURE 2.6. Projection of the individuals over the first principal components.

TABLE 2.6. Percentage of categories inside each cluster for categorical variables.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Global
Dx_HTA	81.3	91.0	88.8	93.9	88.9
Dx_DM	52.1	30.6	30.0	43.4	39.1
Male	53.1	50.5	63.8	48.5	53.4
Female	46.9	49.6	36.3	51.5	46.6
Age_20_65	49.0	27.0	33.8	23.2	33.7
Age_65_75	30.2	30.6	33.8	38.4	33.2
Age_75_100	20.8	42.3	32.5	38.4	33.9
Illiteracy	11.5	17.1	25.0	18.2	17.6
Elementary	29.2	36.9	36.3	39.4	35.5
High School	45.8	24.3	22.5	29.3	30.6
Graduate	13.5	21.6	16.3	13.1	16.3

First, we adjusted models and compared them using cross-validation and the database $N = 386$ patients. Then for the selected method, we fit sequential models. This would give an idea of the utility and reliability of the models. Finally, we uploaded sequential models in a platform that physicians could use as a tool for their daily work.

IIn clinical health, it is common that the predictor variables are related. Then, it is necessary to review the presence of collinearity. In strict terms, collinearity (multicollinearity) appears when a column of X is a linear combination of other columns (Draper & Smith 1998). This is undesirable for the models because it usually leads to unreliable estimates of the regression coefficients. In this case, using package caret (Kuhn et al. 2018), we searched for collinearity and didn't found it.

TABLE 2.7. Mean for quantitative variables for each group.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Global
SysBP	141.2	130.7	131.6	132.5	134.0
DiasBP	77.9	72.9	76.4	74.2	75.2
BMI	26.4	26.0	25.8	26.4	26.2

2.4.1 Selecting Methods Using Cross-Validation

For the $N = 386$ patients, we implemented linear discriminant analysis (Sec. 1.4.1) using the package MASS (Venables & Ripley 2002), logistic regression for multiple categories (Sec 1.4.3) using package nnet (Venables & Ripley 2002), and supervised principal component analysis 1.4.4 (R Core Team 2019).

Using the cross-validation method to compare the classification algorithms (See sec. 1.5 for details), we organized data and, as it could be seen in table 2.8, the model with the best performance is the LDA with the highest percentage of good classification.

TABLE 2.8. Percentage of good classification to compare the three selected classification methods using cross validation with N-folds.

SC Method	Percentage
SPCA	42.8
LDA	84.7
LR	80.8

It should be noted that LDA outperforms the other methods, even in computational time, because cross-validation is implemented internally on the function LDA of package MASS becoming more efficient.

From this moment on, LDA will be the unique model used for the following steps and results.

2.4.2 Sequential Models and Variable Selection

In order to assess the consequences of having less GFR measures, we fit sequential LDA models, that is, to fit them using shorter GFR trajectories. Most of the variables are the same, but the main change remains in the variables associated with trajectories, which are GFR_0 and $GFR_0 - GFR_t$, with $t = 2, 3, \ldots, 15$. In table 2.9 we presented the theoretical structure of all sequential models. Note that all of them have the same number of parameters, a fixed part that does not change between models and the sequential part that depends on the desired length of the trajectory T_t.

In table 2.9 we show the general way to fit all LDA sequential models. There is a base part that is common for all models and some parts that change in between models associated with different GFR measures.

TABLE 2.9. Structure of sequential models.

base	b_1*Dx_DM + b_5*School	+ b_2*Dx_HTA + b_6*BMI	+ b_3*Gender + b_7*SysBP	+ b_4*Age + + b_8*DiasBP
Model_1	base	+ b_9*GFR_0	+ b_{10}*(GFR_1 - GFR_2)	+ b_{11}*(GFR_0 - GFR_2)
Model_2	base	+ b_9*GFR_0	+ b_{10}*(GFR_2 - GFR_3)	+ b_{11}*(GFR_0 - GFR_3)
⋮	⋮	⋮	⋮	
Model_{14}	base	+ b_9*GFR_0	+ b_{10}*(GFR_{14} - GFR_{15})	+ b_{11}*(GFR_0 - GFR_{15})

Again we use the $N = 386$ trajectories and their membership as the response variable, and we fit the fourteen models. Then, using Wilks' lambda to reduce the number of predictor variables (see Sec. 1.5.1 for a brief review and (Kent et al. 2006) for details), we show the structure after variable selection in table 2.10.

For each one of these models, we computed both apparent and cross-validation percentage of good classification (see Sec. 1.5 for details), and the results can be seen in table 2.11. In general, as the number of periods increases, the percentage of good classification increases. Lowest values are in $t = 2$ (cutting at period T_2) and values at apparent percentage are always higher than values from cross-validation, which is very common as the first overestimates the good performance of the model. Additionally, even with reduced variables after selection by Wilks' lambda, the percentage of good classification of the model 14 is the same than in table 2.8.

The last column in table 2.11 is the number of individuals with three or more observations. If we cut trajectories at T_t, it is clear that the number of individuals increases as the number of periods are reduced.

TABLE 2.10. Structure of final sequential models.

base	b_1*Dx_DM	+ b_2*Dx_HTA	+ b_3*Age + b_4*SysBP	+ b_5*DiasBP
Model_1	base	+ b_6*GFR_0	+ b_7*(GFR_1 - GFR_2)	+ b_8*(GFR_0 - GFR_2)
Model_2	base	+ b_6*GFR_0	+ b_7*(GFR_2 - GFR_3)	+ b_8*(GFR_0 - GFR_3)
⋮	⋮	⋮	⋮	
Model_{14}	base	+ b_6*GFR_0	+ b_7*(GFR_{14} - GFR_{15})	+ b_8*(GFR_0 - GFR_{15})

2.4.3 Quality of Models Using the Rest of the Patients

Last models were fitted using $N = 386$ patients. A way to assess the quality of the sequential models is to use the rest: $2445 - 386 = 2069$ patients. The evaluation will be computed as follows:

Step 1: Erase the last GFR measure of the 2069 individuals

Step 2: Predict the class of each one of the 2069 trajectories.

Step 3: Impute the erased measure using the information of the predicted class for each one of the 2069 patients.

TABLE 2.11. Apparent and cross-validation percentages of good classification of sequential models using $N = 386$ patients.

	Apparent (%)	Cross-Val (%).	Non. classif. Ind.
Model_1	59.3	56.4	1543
Model_2	68.9	65.0	1817
Model_3	73.9	72.0	1735
Model_4	76.1	74.5	1580
Model_5	78.6	75.7	1434
Model_6	83.2	81.1	1284
Model_7	87.2	86.1	1140
Model_8	88.2	86.9	960
Model_9	89.0	88.3	817
Model_{10}	88.3	87.2	652
Model_{11}	90.9	88.3	456
Model_{12}	89.6	88.1	293
Model_{13}	88.6	86.3	115
Model_{14}	91.7	89.7	0.0

Additionally, the last column includes non-classified patients (2069) who have more than 3 GFR measures at each model's length.

Step 4: Compare the imputation with the real measure.

For the previous algorithm, in figure 2.7 we show, for a sample of curves, a comparison between imputation and real data. In general, coherent results as imputation (dots) tend to be near real data (line). The difference between dots and triangles is that we can only compare dots against real data, while triangles tend to be future estimations. Also, remembering that for class-prediction the last measure was deleted, results of imputation for the imputations with high bias using the incomplete pattern tend to be coherent, clinically.

As an index of quality of fitting, we computed the Kolmogorov-Smirnov and the chi-squared test (see Massey (1951) and Fisher (1924) for details) between sthe set of patients' last GFR measure and the respective imputation for all the 2069 patients. For the chi-squared test we obtained a p-value of 0.24 and for Kolmogorov-Smirnov a p-value of 0.13. This means, in both cases that, descriptively, both imputation and real data were drawn from the same model.

2.4.4 Sequential Models as a Tool

Using the $N = 386$ patients we could train the sequential models and use them inside a program that graphically helps physicians to track a patient's trajectory and indicate how patients would behave given a set of known variables.

The program is built in R software. For implementation purposes, we decided to use the package **shiny** (Chang et al. 2019) that generates an interactive and aesthetic interface that, in our perspective, will motivate physicians to effectively use our results and that will help them to make wise decisions, even to explain the possible behaviour of the disease for each patient.

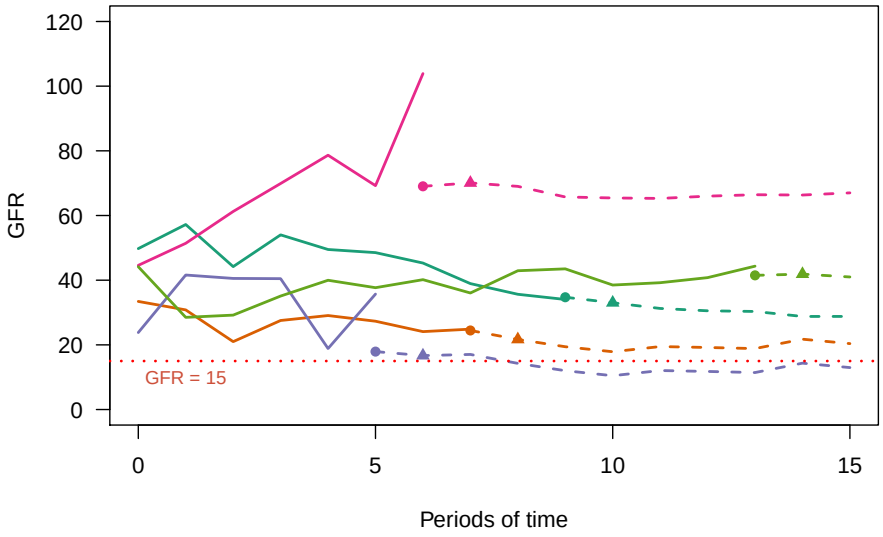


FIGURE 2.7. Comparison of a sample of individuals between imputation (dot symbol) and real curves (continuous lines). The triangle symbol represents the imputation for the next period and dashed lines represent imputation for all possible future periods.

For a patient with at least three GFR measures:

1. The physician introduces information of the clinical variables into the software:
   (a) Diagnostic of diabetes (Dx_DM)
   (b) Diagnostic of hypertension (Dx_HTA)
   (c) Age
   (d) Systolic Tension (SysBP)
   (e) Diastolic Tension (DiasBP)
2. The physician introduces the GFR measures
3. The software shows the next value and the possible future trajectory of the patient for the following periods, using sequential models.

It must be said that the last step is basically an imputation made by $copy - mean$ method (See sec. 1.2) using the prediction of the membership extracted from sequential models. The algorithm is, in essence, quite simple, but very useful for understanding the behavior of CKD via GFR.

Conclusions

- In general, we created a novel strategy to help physicians to make decisions over the treatment of CKD patients by following three steps: Making groups of patients, fitting classification models to understand such partition, and then presenting an estimation of the future trajectory of the patients.
- A first step for a successful research was to conduct a preliminar review and preparation of the data set (sec. 2.2)
- CKD presented high variability between GFR trajectories of the set of individuals. We found that making groups of trajectories is more realistic and convenient. We organized 4 groups of GFR trajectories that captured more than 80% of the total variance, separating fast progressors, low progressors, and two kinds of stable patients. These groups helped making the following results cleaner and more reliable, as we avoided using contrasting phenomenon inside methods (sec. 2.3.2).
- We found for this set of patients, that the LDA model has the best performance, in terms of percentage of good classification against LR and SPCA models (sec. 2.4.1). Also, the LDA model has the lowest computational cost. This coincides with the findings of different authors, like Lebart et al. (1995) and Hastie et al. (2009) who highlighted the good behavior of LDA in real problems.
- We selected the most prominent variables for LDA prediction using backward selection by Wilks' lambda. The set of selected variables were: Dx_DM, Dx_HTA, Age, SysBP, DiasBP, first GFR measure, the difference between the first and the last measure, and the difference between the penultimate and the last GFR measure (sec. 2.4.2).
- Using the selected variables, in order to make the procedure more realistic, we fitted a set of sequential models, taking into account that GFR trajectories could have different sizes. In general, models perform well in terms of percentage of good classification, most of them with a percentage of over 80% (sec. 2.4.2).
- As an approach for future behaviour of individuals, we compared imputations against the last value and we found that such imputation has in general good performance,

giving similar estimations to real data. When imputation is far from real data, it can be associated with extreme changes in GFR patients' trajectories (sec. 2.4.3).

- Using sequential models as theoretical base, we created a graphic interface that can be a useful tool for medical decisions; for it helps the physician to identify, graphically, how patients' CKD is behaving. With more information, it is easier to take more accurate decisions. (see sec. 2.4.4).

- Another advantage of the graphic interface is that the physician would not need to memorize if the first group is of patients with or without diabetes, or if their GFR increases or decreases over time. As the tool does it internally, the physician doesn't have to worry about statistic procedures.

- In all steps we combined clinical expertise and statistical techniques, from the preparation of data (sec. 2.2), to making GFR trajectories' groups 2.3, models' fitting (sec. 2.4) and developing the graphical interface (sec. 2.4.4).

Future Work

- For this investigation, data were usually transversal, but, in case data presented more information, it could be possible to make models that include different types of longitudinal data. For example, if we have proteinuria data longitudinally, both proteinuria and GFR could be used to make groups and not only GFR.
- Unfortunately, for this data set, missing values for some important variables of CKD such as proteinuria or uric acid were high. Then, in case of having this variables, we strongly recommend to assess if they are statistically useful.
- It could be useful to make groups based on values in which recent patient's measures have more weight than those measures made months ago.
- A lack of sensibility was identified when trying to change response variables, that is, group prediction is not strongly modified if a patient has or not diabetes despite the fact that results show that diabetes is an important variable not only statistically, but also medically. Combining these methods with diary clinical expertise and also with more statistical models could lead to a model improvement.
- In this work a classification approach was developed, however, it is possible to make a combination between regression models and classification models by predicting GFR for some periods and then tracking if the prediction helps to improve a patient's classification in the different groups.

References

Atkins, R. C. (2005), 'The changing patterns of chronic kidney disease: the need to develop strategies for prevention relevant to different regions and countries', *Kidney International* **68**, S83–S85.

Barshan, E., Ghodsi, A., Azimifar, Z. & Jahromi, M. Z. (2011), 'Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds', *Pattern Recognition* **44**(7), 1357–1371.

Bradbury, B. D., Fissell, R. B., Albert, J. M., Anthony, M. S., Critchlow, C. W., Pisoni, R. L., Port, F. K. & Gillespie, B. W. (2007), 'Predictors of early mortality among incident us hemodialysis patients in the dialysis outcomes and practice patterns study (dopps)', *Clinical Journal of the American Society of Nephrology* **2**(1), 89–99.

Celeux, G. & Govaert, G. (1992), 'A classification em algorithm for clustering and two stochastic versions', *Computational Statistics & Data Analysis* **14**(3), 315–332.

Chang, W., Cheng, J., Allaire, J., Xie, Y. & McPherson, J. (2019), *shiny: Web Application Framework for R*. R package version 1.3.2.
URL: *https://CRAN.R-project.org/package=shiny*

Chatfield, C., Zidek, J. & Lindsey, J. (2010), *An introduction to generalized linear models*, Chapman and Hall/CRC.

Codreanu, I., Perico, N., Sharma, S. K., Schieppati, A. & Remuzzi, G. (2006), 'Prevention programmes of progressive renal disease in developing nations', *Nephrology* **11**(4), 321–328.

Draper, N. R. & Smith, H. (1998), *Applied regression analysis*, John Wiley & Sons.

Everitt, B. S., Landau, S., Leese, M. & Stahl, D. (2011), *Cluster Analysis, 5th Edition*, John Wiley & Sons, Ltd, Chichester, UK.

Fisher, R. A. (1924), 'The conditions under which χ^2 measures the discrepancy between observation and hypothesis', *Journal of the Royal Statistical Society* pp. 442–450.

Fisher, R. A. (1936), 'The use of multiple measurements in taxonomic problems', *Annals of Eugenics* **7**(2), 179–188.

Genolini, C., Alacoque, X., Sentenac, M., Arnaud, C. et al. (2015), 'kml and kml3d: R packages to cluster longitudinal data', *Journal of Statistical Software* **65**(4), 1–34.

Genolini, C., Écochard, R. & Jacqmin-Gadda, H. (2013), 'Copy mean: a new method to impute intermittent missing values in longitudinal studies', *Open Journal of Statistics* **3**(04).

Hastie, T., Tibshirani, R. & Friedman, J. (2009), 'The elements of statistical learning: Data mining, inference and prediction'.

Hu, B., Gadegbeku, C., Lipkowitz, M. S., Rostand, S., Lewis, J., Wright, J. T., Appel, L. J., Greene, T., Gassman, J., Astor, B. C. et al. (2012), 'Kidney function can improve in patients with hypertensive ckd', *Journal of the American Society of Nephrology* **23**(4), 706–713.

Jolliffe, I. (2002), *Principal component analysis*, Springer Text in Statistics.

Jolliffe, I. T. (1982), 'A note on the use of principal components in regression', *Applied Statistics* pp. 300–303.

Kent, J., Bibby, J. & Mardia, K. (2006), *Multivariate analysis (probability and mathematical statistics)*, Elsevier Amsterdam.

Kuhn, M. C. f. J. W., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer, Z., Kenkel, B., the R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C. & Hunt., T. (2018), *caret: Classification and Regression Training*. R package version 6.0-81.
URL: *https://CRAN.R-project.org/package=caret*

Lebart, L., Morineau, A. & Piron, M. (1995), *Statistique Exploratoire Multidimensionnelle*, Dunond, Paris.

Levey, A. S., Eckardt, K.-U., Tsukamoto, Y., Levin, A., Coresh, J., Rossert, J., Zeeuw, D. D., Hostetter, T. H., Lameire, N. & Eknoyan, G. (2005), 'Definition and classification of chronic kidney disease: a position statement from kidney disease: Improving global outcomes (kdigo)', *Kidney International* **67**(6), 2089–2100.

MacQueen, J. (1967), Some methods for classification and analysis of multivariate observations, *in* 'Proceedings of the fifth Berkeley symposium on mathematical statistics and probability', number 14, Oakland, CA, USA, pp. 281–297.

Massey, F. J. (1951), 'The kolmogorov-smirnov test for goodness of fit', *Journal of the American statistical Association* **46**(253), 68–78.

National Kidney Foundation, . (2017), *Glomerular Filtration Rate (GFR)*. https://www.kidney.org/atoz/content/gfr (accessed August 30, 2018).

Pardo, C. E. & Del Campo, P. C. (2007), 'Combination of Factorial Methods and Cluster Analysis in R: The Package FactoClass', *Revista Colombiana de Estadística* **30**(2), 231–245.

R Core Team (2019), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ripley, B. D. & Hjort, N. (1996), *Pattern recognition and neural networks*, Cambridge University Press.

Suri, R. S., Lindsay, R. M., Bieber, B. A., Pisoni, R. L., Garg, A. X., Austin, P. C., Moist, L. M., Robinson, B. M., Gillespie, B. W., Couchoud, C. G. et al. (2013), 'A multinational cohort study of in-center daily hemodialysis and patient survival', *Kidney International* **83**(2), 300–307.

Todorov, V. (2007), 'Robust selection of variables in linear discriminant analysis', *Statistical Methods and Applications* **15**(3), 395–407.

Venables, W. N. & Ripley, B. D. (2002), *Modern Applied Statistics with S*, fourth edn, Springer, New York. ISBN 0-387-95457-0.

Ward, J. H. (1963), 'Hierarchical grouping to optimize an objective function', *Journal of the American Statistical Association* **58**(301), 236–244.

Warne, R. T. (2014), 'A primerisisisisimo on multivariate analysis of variance (manova) for behavioral scientists.', *Practical Assessment, Research & Evaluation* **19**.

Weihs, C., Ligges, U., Luebke, K. & Raabe, N. (2005), klaR analyzing german business cycles, *in* D. Baier, R. Decker & L. Schmidt-Thieme, eds, 'Data Analysis and Decision Support', Springer-Verlag, Berlin, pp. 335–343.

Wishart, D. (1969), 'An algorithm for hierarchical classifications', *Biometrics* pp. 165–170.