DIPARTIMENTO DI MATEMATICA
"Francesco Brioschi"
POLITECNICO DI MILANO

Multivariate Functional Clustering for the Morphological Analysis of ECG Curves

Ieva, F.;Paganoni,A.M.;Pigoli,D.;Vitelli,V.

Collezione dei *Quaderni di Dipartimento*, numero **QDD 103**
Inserito negli *Archivi Digitali di Dipartimento* in data 24-6-2011

Piazza Leonardo da Vinci, 32 - 20133 Milano (Italy)

Multivariate Functional Clustering for the Morphological Analysis of ECG Curves

Francesca Ieva[♯], Anna Maria Paganoni[♯], Davide Pigoli[♯], Valeria Vitelli[♯]

June 10, 2011

[♯] MOX– Modellistica e Calcolo Scientifico
Dipartimento di Matematica "F. Brioschi"
Politecnico di Milano
via Bonardi 9, 20133 Milano, Italy

francesca.ieva@mail.polimi.it, anna.paganoni@polimi.it,
davide.pigoli@mail.polimi.it, valeria.vitelli@mail.polimi.it

Keywords: ECG signal, Wavelets smoothing, Functional registration, Functional k-means clustering.

Abstract

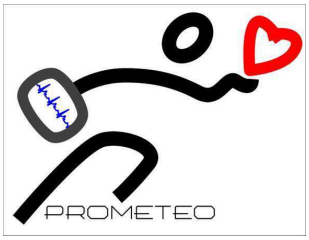
Cardiovascular ischemic diseases are one of the main causes of death all over the world. In this kind of pathologies, it is fundamental to be well-timed in order to obtain good prognosis in reperfusive treatment. In particular, an automatic classification procedure based on statistical analyses of tele-transmitted ECG traces would be very helpful for an early diagnosis. This work presents an analysis on electrocardiographic (ECG) traces (both physiological and pathological ones) of patients whose 12-leads pre-hospital ECG has been sent by life supports to 118 Dispatch Center of Milan. The statistical analysis starts with a preprocessing step, in which functional data are reconstructed from noisy observations and biological variability is removed by a non linear registration procedure. Then, a multivariate functional k-means clustering is carried out on reconstructed and registered ECG curves and their first derivatives. Hence, a new semi-automatic diagnostic procedure, based on the sole ECG's morphology, is proposed to classify ECG traces and the performance of this classification method is evaluated.

1 Introduction

Cardiovascular ischemic diseases are nowadays one of the main causes of death all over the world. In Italy, they are responsible of 44% of overall deaths and call for the most part of emergency rescue operations. In fact, almost all events which require rescue operations to the 118 Milan Dispatch Center (the Italian free toll number for emergencies) concern cardiovascular system. In case of coronary arteries ischemic disease, it

is fundamental to be well-timed in order to obtain good prognosis in reperfusive treatment. This result can be obtained only with pre, inter and intra-hospital networks well organized and synchronized.

Since 2001, a working group collecting 23 Cardiology Units of Milanese urban area and 118 Dispatch Center has been activated. Starting from 2006, this group performs monthly data collection twice a year on all patients admitted to any hospital belonging to the Milan Cardiological Network of Milan with coronary artery disease. The analysis of these data (see Ieva and Paganoni, 2010; Grieco et al., 2007, 2011), pointed out the time of first ECG tele-transmission as the most important factor to guarantee a quick access to an effective treatment for patients (see also Antman et al., 2008).



Then, since 2008, a project named PROMETEO (PROgetto sull'area Milanese Elettrocardiogrammi Teletrasferiti dall' Extra Ospedaliero) has been started with the aim of spreading the intensive use of ECG as pre-hospital diagnostic tool and of constructing a new database of ECGs with features never recorded before in any other data collection on heart diseases. Thanks to the partnerships of Azienda Regionale Emergenza Urgenza (AREU), Abbott Vascular and Mortara Rangoni Europe s.r.l., ECG recorder with GSM transmission have been installed on all Basic Life Supports (BLSs) of Milanese urban area.

In this work we analyse a sample ($n = 198$) of data coming from PROMETEO datawarehouse, which contains all the ECG traces recorded on Milanese urban area by BLSs since the end of 2008. Each file contained in PROMETEO datawarehouse is in correspondence to three sub-files. The first one is called *Details* and contains technical information, useful for signal processing and analysis, such as times of waves' repolarization and depolarization, landmarks indicating onset and offset times of main ECG's subintervals and automatic diagnoses, established by Mortara-Rangoni VERITAS^{TM} algorithm[1]. We used these automatic diagnoses to label ECG traces we analyzed, in order to validate the performances of our unsupervised clustering algorithm. The challenge of this work, in fact, consists of tuning and testing a real time procedure which enables semi automatic diagnosis of the patients' disease based only on ECG traces morphology, then not dependent on clinical evaluations. The second sub-file is called *Rhythm* and contains the ECG signal sampled for 12 seconds (10000 sampled points). The third one is called *Median*. It is built starting from *Rhythm* file, and depicts a *reference* beat lasting 1.2 seconds (1200 points). We carried out the analysis considering only the *Median* files, obtaining 8 curves (one for each ECG lead) for each patient, which represents his/her "Median" beat for that lead.

The main goal of this work is then to identify, from a statistical perspective, specific ECG patterns which could benefit from an early invasive approach. In fact, the identification of statistical tools capable of classifying curves using their shape only could support an early detection of heart failures, not based on usual clinical criteria. To this aim, it is extremely important to understand the link between cardiac physiology and ECG trace shape. As detailed in following sections, we focus on physiological traces in

[1] Mortara Rangoni Europe s.r.l. is the leading provider of ECG algorithms and components for various clinical applications, see http://www.mortara.com.

contrast to Right and Left Bundle Branch Block (RBBB and LBBB respectively) traces. Bundle Branch Block (BBB) is a cardiac conduction abnormality seen on the ECG. In this condition, activation of the left (right) ventricle is delayed, which results in the one ventricle contracting later than the other.

Details on Bundle Branch Blocks and their connection with non-physiological shape of ECG signal will be treated in Section 2, where also clinical details about ECG signals will be given. Wavelet smoothing of ECG traces and their first derivatives and procedure of landmarks registration are explained in Section 3. In Section 4 data analysis is presented, consisting of a multivariate functional k-means clustering of QT-segments of smoothed and registered ECG curves and first derivatives. Finally, in Section 5 results of analysis are discussed, and further developments to be explored in future works are proposed. Alla the analyses are carried out using `R` statistical software (see R Development Core Team, 2009).

2 Electrocardiography and Bundle Branch Block

Electrocardiography is a transthoracic recording of the electrical activity of the heart over time captured and externally recorded through skin electrodes. The ECG works mostly by detecting and amplifying the tiny electrical changes on the skin that are caused when the heart muscle depolarises during each heart beat (for further inquiry about clinical details, see Lindsay, 2006).

First attempts of measuring ECG signals date back to Willem Einthoven (see Einthoven, 1908; Einthoven et al., 1950). The Einthoven *limb leads* (standard leads) are illustrated in Fig. 1 and are defined in the following way:

$$\text{Lead I: } V_I = \Phi_L - \Phi_R, \quad \text{Lead II: } V_{II} = \Phi_F - \Phi_R, \quad \text{Lead III: } V_{III} = \Phi_F - \Phi_L;$$

where

V_I = voltage of Lead I
V_{II} = voltage of Lead II
V_{III} = voltage of Lead III
Φ_L = potential at the left arm
Φ_R = potential at the right arm
Φ_F = potential at the left foot

These lead voltages satisfy the following relationship:

$$V_I + V_{III} = V_{II}, \tag{1}$$

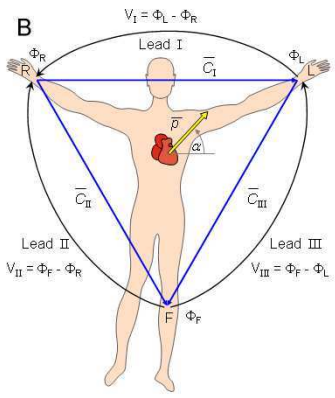


Figure 1: Eithofen limb leads

hence only two of these three leads are independent. A simple model results from assuming that the cardiac sources are represented by a dipole located at the center of a sphere representing the thorax, hence at the center of an equilateral triangle. With these assumptions, the voltages measured by the three limb leads are proportional to the projections of the electric heart vector on the sides of the lead vector triangle. The voltages of the leads are obtained from Equation (1).

Nowadays, the most commonly used clinical ECG-system, the 12-lead ECG system, consists of the following 12 leads: I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, V6. The main reason for recording all 12 leads is that it enhances pattern recognition (see Goldberger 1942a and 1942b; Mason and Likar 1966 and Wilson et al. 1944). Of these 12 leads, the first six are derived from the same three measurement points. Therefore, any two of these six leads include exactly the same information as the other four. So, the ECG traces analyzed in the following sections will consist of leads I, II, V1, V2, V3, V4, V5 and V6 only.

Fig. 2 shows a scheme of the typical shape of a physiological single beat, recorded on ECG graph paper; main relevant points, segments and waves are highlighted. Deflections in this signal are denoted in alphabetic order starting with the letter P, which represents atrial depolarization. The ventricular depolarization causes the QRS complex, and repolarization is responsible for the T-wave. Atrial repolarization occurs during the QRS complex and produces such a low signal amplitude that it cannot be detected, with the exception of physiological ECGs (see Scher and Young, 1957). The direction of travel of the wave of depolarization is named the *heart electrical axis*.

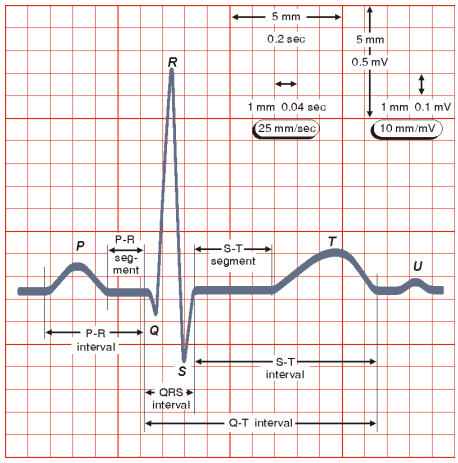


Figure 2: Scheme of the typical shape of a physiological single beat, recorded on ECG graph paper. Main relevant points, segments and waves are highlighted.

In the case of interest, the file *Rhythm* of our dataset represents the output of an ECG recorder. From this curve, a representative heartbeat for each patient is obtained and it is provided in the file *Median*. As we said before, it consists of a trace of a single cardiac cycle (heartbeat), i.e. of a *P* wave, a *QRS* complex, a *T* wave, and a *U* wave, which are normally visible in 50% to 75% of ECGs.

The heart's electrical activity begins in the sinoatrial node (the heart's natural pacemaker, n.1 in Fig. 3), which is situated on the upper right atrium. The impulse travels next through the left and right atria and summates at the AV node (n.2 in Fig. 3). From the AV node the electrical impulse travels down the Bundle of His (n.3 in Fig. 3) and divides into the right and left bundle branches (n.4 and 10 in Fig. 3). The right bundle branch contains one fascicle. The left bundle branch subdivides into two fascicles: the left anterior fascicle and the left posterior fascicle (n.4 and 5 in Fig. 3). Ultimately, the fascicles divide into millions of Purkinje fibres which in turn interdigitise with indi-

vidual cardiac myocytes, allowing for rapid, coordinated, and synchronous physiologic depolarization of the ventricles.

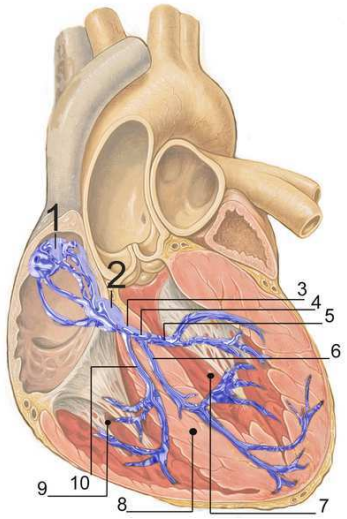


Figure 3: Conduction system of the heart: 1. Sinoatrial node; 2. Atrioventricular node; 3. Bundle of His; 4. Left bundle branch; 5. Left posterior fascicle; 6. Left-anterior fascicle; 7. Left ventricle; 8. Ventricular septum; 9. Right ventricle; 10. Right bundle branch.

Bundle branch or fascicle injuries result in altered pathways for ventricular depolarization. In this case, there is a loss of ventricular synchrony, ventricular depolarization is prolonged, and there may be a corresponding drop in cardiac output.

¿From a clinical perspective, a RBBB typically causes prolongation of the last part of the QRS complex, and may shift the heart electrical axis slightly to the right. LBBB widens the entire QRS, and in most cases shifts the heart electrical axis to the left. Another usual finding with bundle branch block is appropriate T wave discordance: this means that the T wave will be deflected opposite the terminal deflection of the QRS complex.

¿From a statistical point of view, we will focus our analysis on shape modifications induced on the ECG curves and their first derivatives by the BBB pathology, and we will investigate these shape modifications only in a statistical perspective, i.e. not using clinical criteria to classify ECGs. The exploitation of these morphological modifications in the clustering procedure will be the focus of the following Sections.

3 Data smoothing and registration

The dataset coming from PROMETEO datawarehouse consists of the ECG signals of $n = 198$ subjects, among which 101 are Normal and 97 are affected by BBB (49 RBBB and 48 LBBB). As mentioned above, the aim of this work is exploring ECG curves morphology. Thus, the basic statistical unit is the multivariate function which describes heart dynamics, for each patient, on the eight leads.

However, in practice we have only a noisy and discrete observation of the function

describing ECG trace for each patient. Moreover, each patient has his own "biological" time, i.e. the same event of the heart dynamics may happen at different time measurements for different patients: this is only misleading from a morphological point of view. These two problems are common in Functional Data Analysis (FDA) applications and they can be addressed respectively with data smoothing and registration (see Ramsay and Silverman, 2005).

3.1 Wavelets smoothing

The first step of the statistical analysis consists in data smoothing starting from noisy measurements: to this aim, the choice of the functional basis is crucial. Wavelet bases seem suitable for our data because every basis function is localized both in time and in frequency, being therefore able to capture ECG strong localized features (peaks, oscillations...). In particular we use a Daubechies wavelet basis with 10 vanishing moments (see Daubechies, 1988 for details), because we are interested also in derivatives of the ECG traces and thus we need a basis smooth enough for this purpose.

As in most smoothing methods based on wavelet expansion, it is necessary to deal with a grid of 2^J points, $J \in \mathbb{N}$. Thus, in the further analysis we use only the central $2^{10} = 1024$ observation points. There is no loss of significant information: the portion of the signal on which we focus the analysis contains all the important features of the ECG trace. For this reason, we choose not to turn to non-decimated wavelets, which could be applied also to non dyadic grid but require a larger computational effort.

Since the eight leads (i.e. I, II, V1, V2, V3, V4, V5 and V6) jointly describe the complex heart dynamic, when smoothing these data it is appropriate to use a technique which takes into account all the eight leads simultaneously. This helps in detecting significant features, which reflect on more then one leads. To this aim in Pigoli and Sangalli (2011) it is developed a wavelet based smoothing technique for multivariate curves. This technique is used to obtain the estimation of 8 dimensional ECG signals. It has also the advantage to provide an estimate of derivatives, which is straightforward when the estimate is provided in functional basis expansion: it can be obtained simply by a linear combination of the basis functions derivatives.

Thus, starting from the vectorial raw signal, we estimate the vectorial function

$$\mathbf{f}_i(t) = (I_i(t), II_i(t), V1_i(t), V2_i(t), V3_i(t), V4_i(t), V5(t), V6(t)),$$

and its derivatives, for each patient $i = 1, \ldots, 198$. See Pigoli and Sangalli (2011) for a detailed description of this smoothing procedure. Fig. 4 shows raw data and functional estimates obtained with this wavelet smoothing procedure for a normal subject. Observations are now in a functional form and thus we can use FDA techniques. The smoothing procedure is essential also for an accurate derivative reconstruction, as shown in Fig. 5, where the estimate of the first derivative is superimposed to the first central finite difference (i.e. a rough indication of first derivative behavior).

3.2 Landmark registration

Functional observations usually show both phase and amplitude variation, i.e. each curve has its own biological time so that the same feature can appear at different times among the patient. It is well known that a correct separation between these two kind

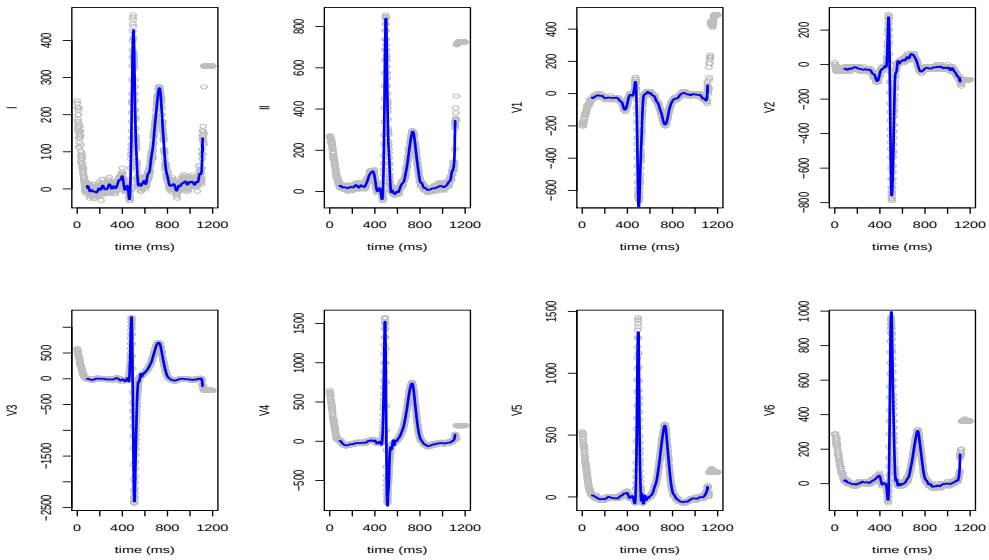


Figure 4: Raw data of the eight leads (black points) and wavelet functional estimates (blue) for a normal subject.

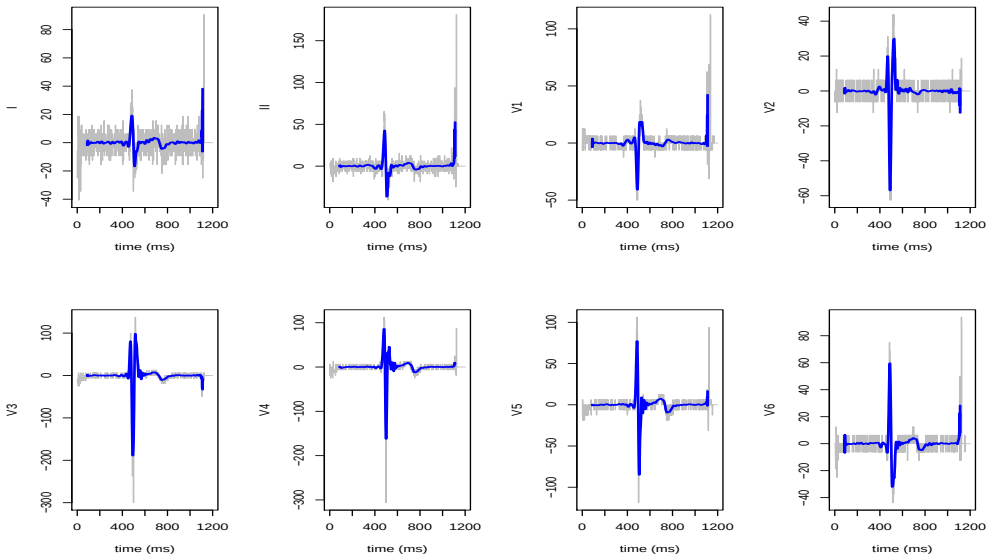


Figure 5: First central finite difference of the eight leads (gray) and wavelet estimates of the first derivatives (blue) for a normal subject.

of variability is necessary for a successful analysis (see Ramsay and Silverman, 2005). We address this problem through a registration procedure based on landmarks, which are points of the curve that can be associated with a specific biological time. Five of these landmarks are provided by Mortara-Rangoni procedure and can be found in the *Details* file. They identify the P wave (P_{onset}, P_{offset}), QRS complex (QRS_{onset}, QRS_{offset})

Table 1: Landmarks obtained at the end of the registration procedure, as the mean of landmarks of all the curves, and used to select the portion of smoothed and registered ECG curves relevant to our analysis (first line of the table); in the second line, landmarks standard deviations. Landmarks values are referred to a registered time in ms.

	P^0_{onset}	P^0_{offset}	QRS^0_{onset}	$Ipeak^0$	QRS^0_{offset}	T^0_{offset}
mean	184.3	298.2	354.8	407.2	476.9	755.8
standard deviation	39.7	37.4	18.9	15.4	21.4	44.2

and T wave (T_{offset}). We add one more landmark corresponding with the R peak on the I lead ($Ipeak$). We choose this landmark because only on the I lead both physilogical and pathological ECG traces present a clearly identifiable R peak. Since all the leads capture the same heart dynamics, biological time must be the same. Thus, these landmarks can be used to register all the leads. For each patient i we look for a warping function h_i such that

$$h_i(P_{onset}) = P^0_{onset} \qquad h_i(P_{offset}) = P^0_{offset}$$

$$h_i(QRS_{onset}) = QRS^0_{onset} \quad h_i(Ipeak) = Ipeak^0$$

$$h_i(QRS_{offset}) = QRS^0_{offset} \quad h_i(T_{offset}) = T^0_{offset}$$

where P^0_{onset}, P^0_{offset}, QRS^0_{onset}, $Ipeak^0$, QRS^0_{offset} and T^0_{offset} are the mean values of the correspondent landmarks. These values are reported in Table 1, together with the associated standard deviations. We solve this problem using spline interpolation of degree 3. Thus, the registered vectorial function will be

$$\mathbf{F}_i(t) = \mathbf{f}_i(h_i(t)),$$

for every patient $i = 1, \ldots, 198$. Fig. 6 shows both unregistered and registered I leads for all the 198 patients. This is a non linear registration procedure, since in this framework there is no simple affine transformation which can take in account the subject specific variability. The registration procedure separates morphological information (i.e. amplitude variability) from duration of the different segments of ECG (i.e. phase variability). The former is captured by the registered ECG traces, while the latter is described by warping functions, determined by landmarks. In clinical practice the duration of different segments of ECG and particularly the QRS complex length is one of the most important parameters to identify pathological situations. However, this kind of information is not able to distinguish among different pathologies, such as Right and Left BBB. This can be seen also in our dataset. If we perform a multivariate 3-means algorithm on interval lengths ($P_{offset} - P_{onset}$, $QRS_{onset} - P_{offset}$, $QRS_{offset} - QRS_{onset}$ and $T_{offset} - QRS_{offset}$), with the aim of identifying the existing 3 groups, we obtain the result shown in Table 2: this method correctly separates physiological traces from pathological ones but it gives no information on the pathology. For this reason, we focus our analysis on the registered curves, in the attempt to extract other diagnostic information from ECG morphology. In clinical practice, the result of our analysis should be considered together with traditional diagnostic tools based on segment lengths.

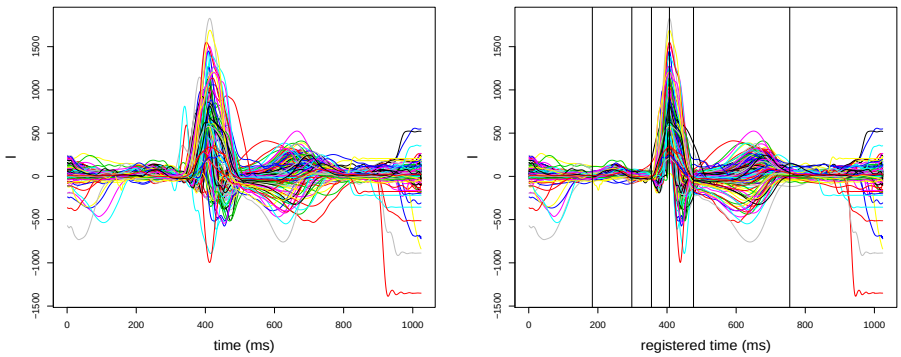


Figure 6: Original I leads for the 198 patients (left) and registered ones (right). Vertical lines indicate position of mean landmarks P^0_{onset}, P^0_{offset}, QRS^0_{onset}, $Ipeak^0$, QRS^0_{offset}, T^0_{offset} .

Table 2: Confusion matrix related to patients disease classification. Results are obtained performing 3-means clustering algorithm on interval lengths.

	Normal	RBBB	LBBB
Cluster 1	96	6	0
Cluster 2	2	17	25
Cluster 3	3	26	23

4 Data analysis

In this section we propose the use of FDA techniques to perform clustering of smoothed and registered ECG traces. Aim of the analysis is the development of a proper classification procedure, able to distinguish the grouping structure induced in the sample of ECGs by the presence of different pathologies, on the basis of the sole shape of the considered curves.

As previously discussed in Section 2, ECG traces are very complex functional data, in which different portions of the domain can be analyzed in order to detect different pathologies. The main focus of our analysis stands in the investigation of BBB pathology, which mainly expresses in the ECG trace through a lengthening of the QRS complex and a modification of the T wave. In fact, the diagnosis of BBB is not concerned with modifications in P wave, since this portion of the ECG curve deals with cardiac rhythm dysfunctions our patients are not affected by. We thus focus our classification analysis on the QT-segment. Since we have already registered the ECG signals, all the curves show relevant features at the same time points, corresponding to the reference landmarks P^0_{onset}, P^0_{offset}, QRS^0_{onset}, $Ipeak^0$, QRS^0_{offset}, T^0_{offset} (see Section 3.2): this fact allows us to select, for all the registered curves of the dataset, only the portion of ECG trace belonging to the interval $[P^0_{offset}, T^0_{offset}]$, which is relevant to our diagnostic purposes.

In particular, we select only the portion of

$$\mathbf{F}(t) = \{F^r(t)\}_{r=1}^8 = (I(t),\, II(t),\, V1(t),\, V2(t),\, V3(t),\, V4(t),\, V5(t),\, V6(t))$$

such that $t \in T := [P^0_{offset}, T^0_{offset}]$, where P^0_{offset} and T^0_{offset} are the values reported in the first line, second and sixth columns of Table 1.

4.1 Functional classification

We analyze the n patients according to a functional k-means clustering procedure, in which all the eight leads $\mathbf{F}_i(t) : T \to \mathbb{R}^8$, for patients $i = 1,\ldots,n$, are simultaneously clustered. To develop this clustering procedure we suppose that $\mathbf{F}_i(t) \in H^1(T;\mathbb{R}^8)$. Since we consider all the eight leads simultaneously in the analysis, we name the employed clustering procedure *multivariate functional k-means*, to distinguish it from *standard functional k-means*, which would treat each lead separately.

A proper definition of functional k-means procedure and an introduction to its consistency properties can be found in Tarpey and Kinateder (2003). We develop a similar k-means procedure, choosing the following distance between ECG traces

$$d_1(\mathbf{F}_i(t),\mathbf{F}_j(t)) = \sqrt{\sum_{r=1}^{8} \int_T (F_i^r(t) - F_j^r(t))^2 dt + \int_T (DF_i^r(t) - DF_j^r(t))^2 dt}, \qquad (2)$$

for $i, j = 1,\ldots,n$, and with $DF_i^r(t)$ being the wavelet estimate of the first derivative of the r-th lead in the ECG trace of the i-th patient. Note that the distance defined in (2) is the natural distance in the Hilbert space $H^1(T;\mathbb{R}^8)$.

In order to perform comparisons, and to test the robustness of our clustering procedure, we considered two more distances between two ECG traces

$$\tilde{d_1}(\mathbf{F}_i(t),\mathbf{F}_j(t)) = \sqrt{\sum_{r=1}^{8} \int_T (DF_i^r(t) - DF_j^r(t))^2 dt}, \qquad (3)$$

$$d_2(\mathbf{F}_i(t),\mathbf{F}_j(t)) = \sqrt{\sum_{r=1}^{8} \int_T (F_i^r(t) - F_j^r(t))^2 dt}. \qquad (4)$$

The distance defined by (3) is the natural semi–norm in the Hilbert space $H^1(T;\mathbb{R}^8)$, while the one defined in (4) is the norm in the Hilbert space $L^2(T;\mathbb{R}^8)$: they are both considered in the clustering procedure not only to compare performances of multivariate functional k-means under different specifications of the distance, but also to have an insight on the role of curves first derivatives: we claim that both the ECG trace and its first derivative are essential to distinguish more similar morphologies from less similar ones.

Functional k-means clustering algorithm is an iterative procedure, which alternates a step of *cluster assignment*, in which all curves are assigned to a cluster, and a step of *centroid calculation*, in which a relevant functional representative (the centroid) for each cluster is identified.

More precisely, in the cluster assignment step each curve is assigned to the cluster whose centroid (computed at the previous iteration) is nearer according to the distances

defined in (2), (3) or (4) respectively. Instead, the identification of centroids $\varphi_l(t)$ for $l = 1,\dots,k$, is performed solving the following optimization problem

$$\varphi_l(t) = \underset{\varphi\in\Omega_d}{\operatorname{argmin}} \sum_{i:C_i=l} d(\mathbf{F}_i(t),\varphi(t))^2,$$

where C_i is the cluster assignment of the i^{th} patient at the current iteration, d is one of the three distances defined in (2-4), and Ω_d is the Hilbert space with respect to which the chosen distance d is natural. The solution to this infinite dimensional optimization problem obviously depends on the choice of the distance: it is possible to prove that, both when the distance is measured with (2), and when it is measured with (4), the minimizer $\varphi_l(t)$ corresponds to the functional mean of curves belonging to the same cluster. An immediate consequence of this result is that, when the semi–norm in H^1 (eq. (3)) is used, the centroid is the functional mean of the first derivatives of curves belonging to the same cluster.

There are many different implementations of functional k-means algorithm in the literature on functional data analysis, among which some procedures integrate registration in the classification steps (e.g. the k-means alignment algorithm described in Sangalli et al., (2010), the core shape modeling approach in Boudaoud et al., (2010), the non-parametric time–synchronized iterative mean updating technique in Liu and Müller, (2003), or finally the SACK model in Liu and Yang, (2009)). Here, instead, we chose to separate registration and clustering in two subsequent steps of the analysis, since the latter doesn't use any information beside morphology of the ECG traces, while the former is based on a strong clinical indication provided by landmarks supplied by the Mortara-Rangoni VERITAS$^{\text{TM}}$ algorithm.

The k-means clustering procedure clearly depends not only on the choice of the distance, but also on the number of clusters k. Being the number of clusters a–priori unknown, we also consider a way to select the optimal number of clusters k^* via silhouette values and plot of the final classification (see Struyf et al., 1997). In particular, the silhouette plot of a final classification consists in a bar plot of the *silhouette values* s_i, obtained for each patient $i = 1,\dots,n$ as

$$s_i = \frac{b_i - a_i}{max\{a_i, b_i\}},$$

where a_i is the average distance, according to one of the three distances defined in (2-4), of the i^{th} patient to all other patients assigned to the same cluster, while

$$b_i := \min_{l=1,\dots,k;l\neq C_i} \frac{\sum_{j:C_j=l} d(\mathbf{F}_i(t),\mathbf{F}_j(t))}{\#\{j : C_j = l\}}$$

is the minimum average distance of the i^{th} patient from another cluster, where d is one of the three distances defined in (2-4). Clearly s_i always lies between -1 and 1, the former value indicating a misclassified patient, while the latter a very well classified one. Note that a patient which alone constitutes a cluster, has silhouette value equal to 1, but he is not considered in the silhouette plot for choosing k^*.

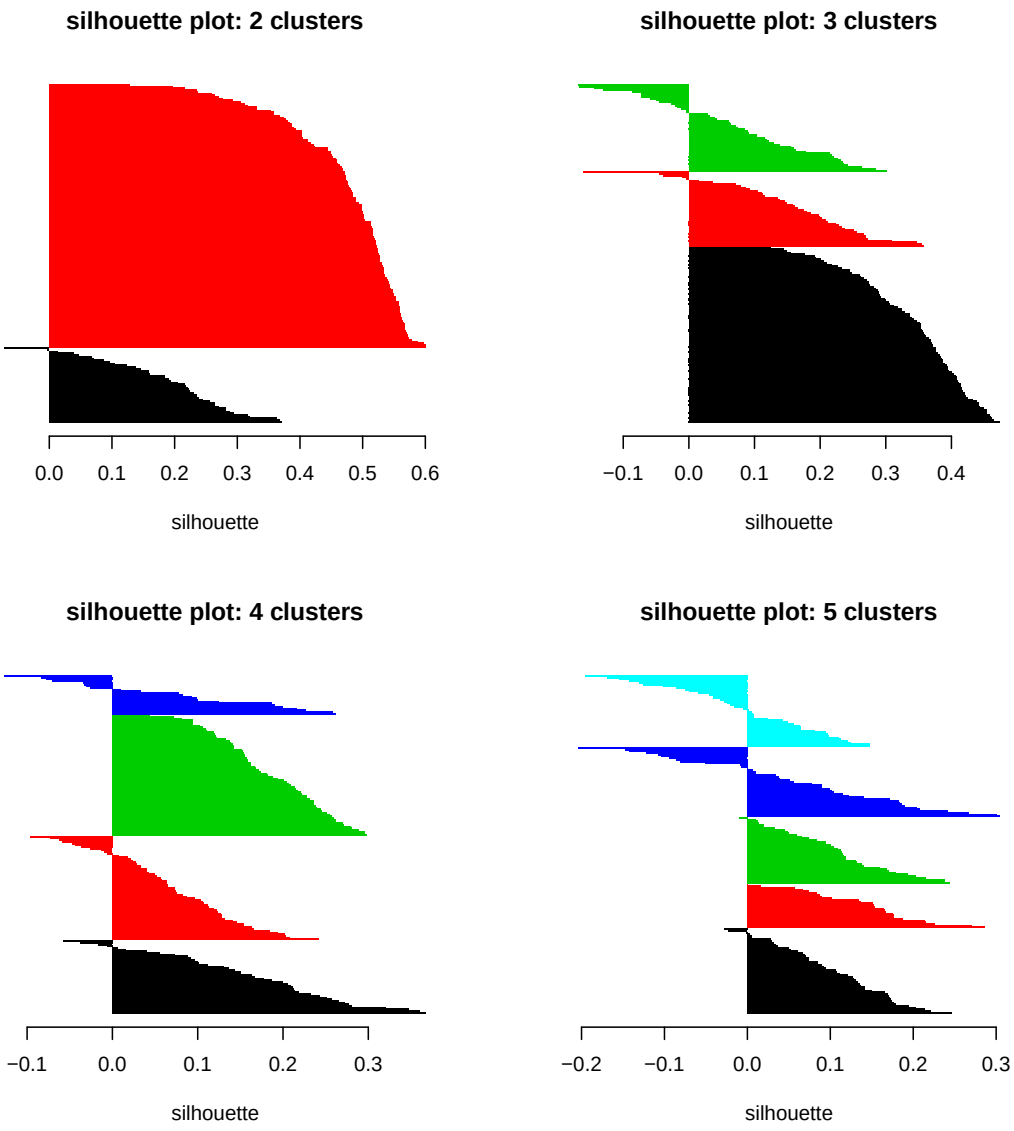


Figure 7: Silhouette plots of the clustering result obtained via multivariate functional k-means procedure, setting $k = 2,3,4,5$ and with distance given by (2); data are ordered according to an increasing value of silhouette within each cluster, and are coloured according to the cluster assignment.

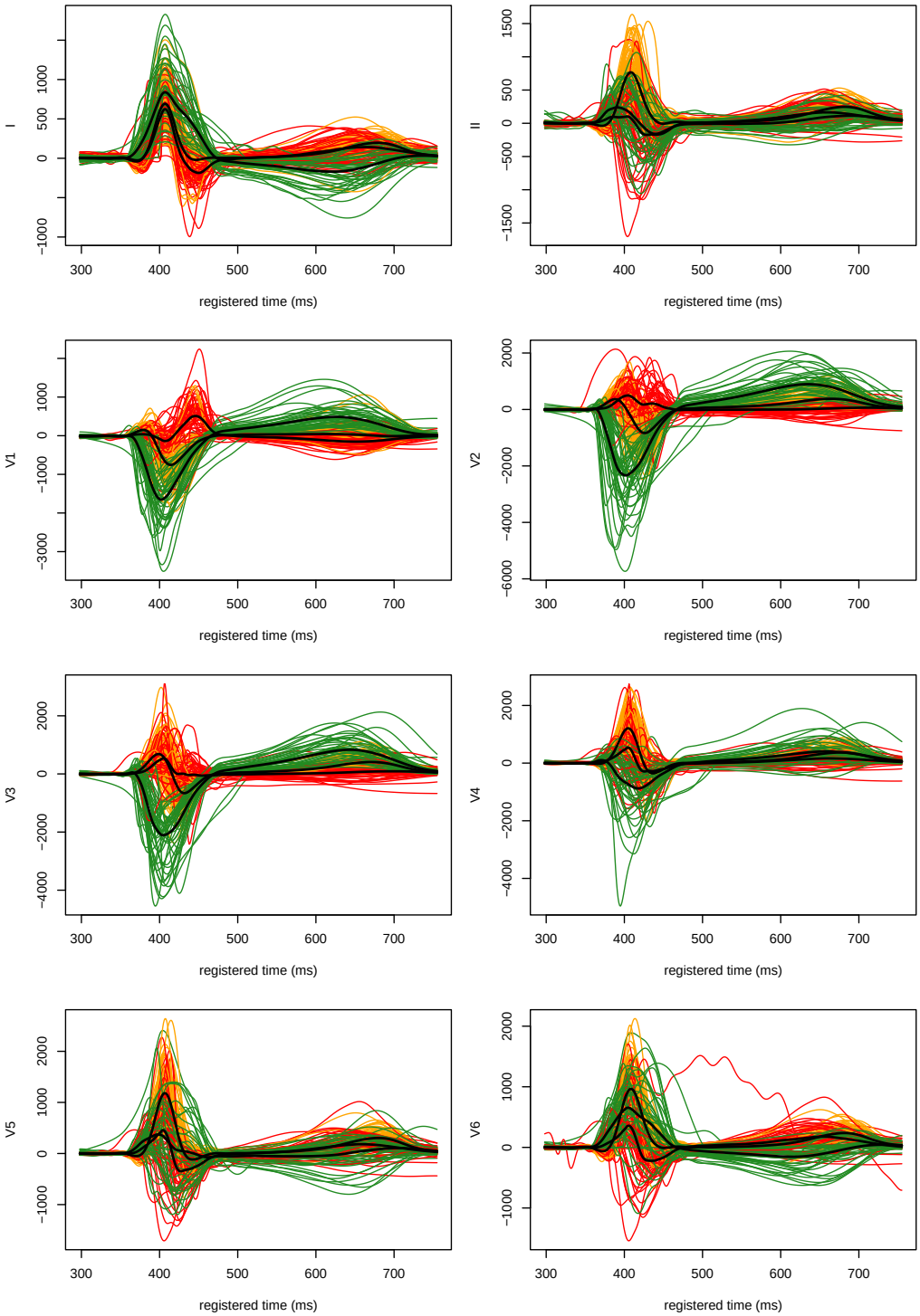


Figure 8: Smoothed and registered ECG traces (QT-segment): the whole dataset is coloured according to the final cluster assignments of multivariate functional 3-mean clustering, with distance given by (2); the superimposed black lines are the three final cluster centroids (functional means). Each panel correspond to a different lead of the ECG traces.

4.2 Results and discussion

Aim of the analysis is to detect the underlying grouping structure in our sample of 198 ECG traces. We thus perform clustering of the whole dataset via the multivariate functional k-means algorithm previously described, using the different definitions of the distance between curves given in (2-4).

The final silhouette plots obtained by clustering the sample of 198 ECG traces according to a multivariate functional k-means procedure with distance d_1 (2), and setting $k = 2,3,4,5$, are shown in Fig. 7. As we can appreciate from the picture, the grouping structure obtained setting $k = 3$ seems the best one, both in terms of silhouette profile, and in terms of wrong assignments. A similar result is obtained measuring the distance between curves via (3) or (4); however, the procedure seems to detect the best grouping structure when both the curves and their derivatives are considered in the distance. We thus set $k^* = 3$.

The final classification obtained with this choice of the distance, and setting $k = 3$, is shown in Fig. 8, where the whole functional dataset is coloured according to cluster assignments; each panel corresponds to a different lead. From inspection of this picture a different shape of ECGs assigned to different clusters can be immediately appreciated, especially looking at the final centroids (functional mean) of each group, drawn in black in each panel of the picture. We shall now verify whether this difference in the ECGs morphology across clusters is due to the different pathology.

Since we have an indication of the different pathologies of the patients included in the sample, we can analyze the confusion matrix associated to the final cluster assignments, with respect to the Mortara-Rangoni algorithm classification (Normal, RBBB and LBBB). The confusion matrices obtained via multivariate functional k-means with different choices of the distance between curves (given by d_1, $\tilde{d}_1$ or d_2) are shown in Table 3. We remark that the final cluster assignments are based on the sole shape of the smoothed and registered ECG curves and their first derivatives, analyzed via a unsupervised classification procedure.

Both choosing the H^1 norm and the L^2 norm, the results seem appreciable, and slightly better in the former case: the final grouping structure traces out quite coherently the patients disease classification, with only few cases wrongly assigned. Moreover, we remark the improvement in the results obtained via multivariate functional 3-means with respect to the results of 3-means clustering algorithm on interval lengths (see Table 2): we are now able not only to detect pathological subjects, but also to distinguish between the two different pathologies present in the dataset. The result obtained via multivariate functional 3-means clustering with H^1 semi-norm, instead, is not so positive, since cluster 1 and 2 apparently merge physilogical traces with ECGs of patients affected by RBBB.

The effectiveness of the clustering procedure in detecting the grouping structure among data suggests the definition of a semi–automatic diagnostic procedure based on the multivariate functional k-means algorithm: in fact, the final result of our clustering procedure is a set of k centroids, representative of each cluster, which can be used as reference signals to compare a new ECG trace. Suppose a new ECG signal is available: we could have an immediate hint on the new patient's diagnosis by smoothing its ECG trace, registering it and finally assigning it to the group characterized by the nearest centroid.

Table 3: Confusion matrices related to patients disease classification. Results are obtained by application of multivariate functional 3-means clustering algorithm to smoothed and registered QT-segment of ECG curves, with different choices of the distance between ECGs: H^1 norm (eq. (2), first table), H^1 semi-norm (eq. (3), second table) and L^2 norm (eq. (4), third table). In the first table, cluster 1,2,3 respectively correspond to orange, green and red in Fig. 8.

	Normal	RBBB	LBBB
1	95	7	1
2	6	42	3
3	0	0	44

	Normal	RBBB	LBBB
1	71	12	0
2	30	36	5
3	0	1	43

	Normal	RBBB	LBBB
1	94	6	2
2	7	43	3
3	0	0	43

It is important to evaluate the *misclassification cost* for this procedure, with the choice of the different functional distances. To this aim, we perform a *cross-validation analysis*. We randomly choose among ECGs a training set of 80 Normal subjects, 40 to RBBBs and 40 to LBBBs, for a total of $n_{training} = 160$ curves. A multivariate functional 3-means clustering is performed on the selected training set; we then consider the remaining $n_{test} = 38$ curves, and we assign each of them to the cluster whose centroid is nearer, according to distances (2-4). Given the patients disease classification, we compute misclassification cost using the following index

$$cost_{CV} = \frac{\lambda_1 \cdot misc_N + \lambda_2 \cdot (misc_{RN} + misc_{LN}) + \lambda_3 \cdot (misc_{RL} + misc_{LR})}{n_{test}}, \qquad (5)$$

where $misc_N$ is the number of healthy patients assigned to a pathological cluster[2], $misc_{RN}$ and $misc_{LN}$ are the number of patients respectively affected by RBBB and LBBB which are assigned to the cluster of healthy patients, while $misc_{RL}$ and $misc_{LR}$ are the number of patients whose ECGs are detected as pathological, but whose pathology is wrong. The parameters λ_1, λ_2 and λ_3 are misclassification weights: they are chosen according to the suggestion of the clinicians, who believe that assigning a BBB patient to the cluster of healthy patients is approximately 4 times more serious than treating as pathological a normal subject, which indeed is two times more serious than assigning a RBBB patient to the LBBB cluster (or viceversa); in order to determine the values of the weights we introduce a further request: $cost_{CV} = 1$ in the worst case, i.e. when all

[2] given the final cluster assignments, the cluster of healthy patients is detected as the one that includes the most physiological traces. The pathological ones are subsequently chosen, first the one that contains the more RBBB traces, while the cluster that remains is the LBBB one.

Table 4: Mean misclassification cost (first row) and standard deviation (second row) computed over 20 repetitions of the cross-validation procedure via eq. (5).

distance	d_1	$\tilde{d}_1$	d_2
mean $cost_{CV}$	0.1227563	0.2286588	0.1275316
std dev $cost_{CV}$	0.1112663	0.1050911	0.1220574

Normal subjects are classified as BBB and all BBB subjects are classified as Normal. This led to the choices $\lambda_1 = 0.4270$, $\lambda_2 = 1.7079$ and $\lambda_3 = 0.2135$.

We repeat this procedure 20 times, computing each time the misclassification cost according to eq. (5): the mean and standard deviation computed along the 20 cross-validation repetitions are shown in Table 4. Even if all the distances (2-4) provide good results, we notice that the norm in the Hilbert space $H^1(T;\mathbb{R}^8)$ seems to give best results, thus confirming our initial claim: both registered curves and first derivatives are needed to accurately compare ECGs morphology.

5 Conclusions

In this work we proposed a statistical framework for analysis and classification of ECG curves starting from their sole morphology. We analyzed a database composed by 198 ECG traces - 101 of them were Normal, 49 were RBBB and 48 were LBBB - extracted from PROMETEO datawarehouse. The strongly localized features (peaks, oscillations...) of ECG curves makes them particulary suited to be smoothed via wavelets methods, since every basis function is localized both in time and in frequency; to this aim, and to reconstruct smoothed curves together with their first derivatives, we used a Daubechies wavelet basis with 10 vanishing moments. Moreover, being ECGs functional observations, they show both phase and amplitude variation, i.e. the same features can appear at different times among the patients. Since a correct separation between these two kind of variability is necessary for a successful analysis, we register ECG traces, choosing a landmark based procedure, which identifies as landmarks those time points that can be associated with a specific biological event. Five of them are provided by the Mortara-Rangoni VERITAS^{TM} algorithm, identifying the P wave, the QRS complex and the T wave; we add one more landmark corresponding to the peak of R wave on the I lead, an easily localized feature on each ECG. In this way, we managed to separate morphological information of the curves (i.e. amplitude variability) from duration of each ECG interval (i.e. phase variability).

We chose to analyze morphological information via a multivariate functional k-means, thus simultaneously clustering all 8 leads of each patient, with three different choices for the distance between ECGs, involving curves and/or first derivatives; our claim is that both the ECG trace and its first derivative are necessary to deeply capture

the morphological characteristics of ECGs. The optimal number of clusters can be chosen via a measure of the goodness of the clustering results, and in all considered cases it is set equal to 3. The confusion matrix resulting from our classification framework shows appreciable results, especially when the distance considers both curves and first derivatives, confirming our initial claim. Thus, we propose a classification procedure which uses groups centroid as reference signals. This technique could help in the semi–automatic diagnosis of BBB–related pathologies. We perform a cross-validation analysis to evaluate the misclassification cost associated to this procedure: our algorithm performances seem very appreciable, especially when functional distance considers both ECG curves and their first derivatives. The proposed classification procedure has an extreme generality, due to the flexibility in the definition of distance between functional data.

In fact, the innovative aspect of this proposal consists in developing advanced statistical methods aimed at detecting pathological ECG traces (in particular Bundle Branch Blocks), starting only from morphological features of the curves. This allows for diagnoses consistent with clinical practice, starting from purely statistical considerations.

Further refinements of our clustering procedure could help in its integration in the cardiovascular context, possibly for the diagnosis of different kinds of pathologies (not only BBB); due to the extreme generality of the algorithm, which is based only on morphological characteristics of the curves, this generalization can be based on a proper definition of a distance between functional data, e.g. including higher order derivatives.

Acknowledgment

This work is within PROMETEO (PROgetto sull'area Milanese Elettrocardiogrammi Teletrasferiti dall'Extra Ospedaliero). The authors wish to thank in particular dr. Niccolò Grieco for clinical counselling and Ing. Johan DeBie for technical support. Finally, thanks to AREU, 118 Milan Dispatch Center and Mortara Rangoni Europe s.r.l. for having provided data.

References

[1] Antman, E.M., Hand, M., Amstrong, P.W., Bates, E.R., Green L.A. & al. (2008) Update of the ACC/AHA 2004 Guidelines for the Management of Patients with ST Elevation Myocardial Infarction, *Circulation*, 117, 269-329.

[2] Boudaoud, S., Rix, H., and Meste, O. (2010), Core Shape modelling of a set of curves, *Computational Statistics and Data Analysis*, 54, 308–325.

[3] Daubechies, I. (1988), Orthonormal basis of compactly supported wavelets, *Communictions on Pure and Applied Mathematics*, **41**, 909–996.

[4] Einthoven, W. (1908), Weiteres ber das Elektrokardiogram, *Pflger Arch. ges. Physiol.*, 122, 517–48.

[5] Einthoven, W., Fahr, G. and de Waart, A. (1950), On the direction and manifest size of the variations of potential in the human heart and on the influence of

the position of the heart on the form of the electrocardiogram. *American Heart Journal*, 40(2), 163–211.

[6] Goldberger, E. (1942a), The aVL, aVR, and aVF leads: a simplification of standard lead electrocardiography. *American Heart Journal*, 24, 378–96.

[7] Goldberger, E. (1942b), A simple indifferent electrocardiographic electrode of zero potential and a technique of obtaining augmented, unipolar extremity leads. *American Heart Journal*, 23, 483–92.

[8] Grieco, N., Sesana, G., Corrada, E., Ieva, F., Paganoni, A.M., Marzegalli M. (2007). The Milano Network for Acute Coronary Syndromes and Emergency Services. *MESPE journal*, First Special Issue 2007.

[9] Grieco, N., Ieva, F., Paganoni, A.M. (2011). Performance assessment using mixed effects models: a case study on coronary patient care. *IMA Journal of Management Mathematics*, in press.

[10] Ieva, F., Paganoni, A.M. (2010). Multilevel models for clinical registers concerning STEMI patients in a complex urban reality: a statistical analysis of MOMI^2 survey, *Communications in Applied and Industrial Mathematics*, 1 (1), 128 - 147.

[11] Lindsay, A.E. (2006), ECG learning centre, [online] http://library.med.utah.edu/kw/ecg/index.html

[12] Liu, X., and Müller, H.-G. (2003), Modes and clustering for time–warped gene expression profile data *Bioinformatics*, 19(15), 1937–1944.

[13] Liu, X., and Yang, M. (2009), Simultaneous curve registration and clustering for functional data *Computational Statistics and Data Analysis*, 53, 1361–1376.

[14] Mason, R., Likar, L. (1966), A new system of multiple leads exercise electrocardiography, *American Heart Journal*, 71(2), 196–205.

[15] Pigoli, D. and Sangalli, L.M. (2011), "Wavelets in Functional Data Analysis: estimation of multidimensional curves and their derivatives", Tech. Rep. MOX 09/2011, Dipartimento di Matematica, Politecnico di Milano.
[online] http://mox.polimi.it/it/progetti/pubblicazioni/quaderni/09-2011.pdf

[16] R Development Core Team (2009), "R: A language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria. [online] http://www.R-project.org.

[17] Ramsay, J.O. and Silverman, B.W. (2005), *Functional Data Analysis* (2nd ed.), Springer, New York.

[18] Sangalli, L.M., Secchi, P., Vantini, S., and Vitelli, V. (2010), *k*-mean alignment for curve clustering, *Computational Statistics and Data Analysis*, 54, 1219–1233.

[19] Scher, A.M. and Young, A.C. (1957), Ventricular depolarization and the genesis of the QRS, *Annals of New York Academy of Science*, 65, 768–78.

[20] Struyf, A., Hubert, M., and Rousseeuw, P. (1997), Clustering in an Object-Oriented Environment, *Journal of Statistical Software*, 1, 4, 1–30.

[21] Wilson, F.N., Johnston, F.D., Rosenbaum, F.F., Erlanger, H., Kossmann, C.E., Hecht, H., Cotrim, N., Menezes de Olivieira, R., Scarsi, R., Barker, P.S. (1944), The precordial electrocardiogram, *American Heart Journal*, 27, 19–85.

[22] Tarpey, T., and Kinateder, K. K. J. (2003), Clustering Functional Data, *Journal of Classification*, 20, 93–114.