Characterizing the Spatial Distribution of Geolocated Categorical Values

Pedro J. Zufiria[1*], Miguel Á. Hernández-Medina[2]

[1] Dep. Matemática Aplicada a las TIC, Information Processing and Telecommunications Center (IPTC).
[2] ETSI Telecomunicación, Universidad Politécnica de Madrid, Spain.

* Corresponding author. Tel.: +34 910672286; email: pedro.zufiria@upm.es

Manuscript submitted May 14, 2018; accepted June 1, 2018.
doi: 10.17706/ijapm.2019.9.1.47-53

Abstract: We analyze the existence of some regularities in the spatial distribution of labelled individuals. Several indexes for measuring the relationship between the non-ordered categorical variable and the geolocation variables are evaluated together with a new proposed one. Simulations suggest that this new index is quite robust and efficient when compared with the previously known ones.

Key words: Geolocated date, correlation analysis, categorical variables, voronoi tessellation, entropy.

1. Introduction

The increasing availability of large data sets containing heterogeneous information [1] encourages the development of new tools for the statistical analysis of random vectors containing different types of variables such as continuous, discrete and categorical ones. Existing techniques to characterize the relationship between random variables usually consider that these variables are of the same type. For instance, classical Pearson's and G-test [2], [3] assess the independence between discrete variables; on the other hand, the independence between continuous variables has been addressed via several ways such as binning techniques [4], mutual information estimators [5], [6], kernel based methods [7], correlation distance estimators [8] or detectors based on the analysis of subsequences [9]. Regarding the relationship between heterogeneous variables, it can also be assessed using some of the above procedures after a previous binning step; in addition, ANOVA-type tests can also be employed in some specific scenarios. Alternatively, some new tools have already been developed [10] which address the estimation of the mutual information between discrete and continuous variables. Most of the mentioned techniques behave well in specific scenarios, whereas the analysis of their behavior in other scenarios remains a challenging problem.

Many sources of data such as Call Detail Records (CDRs) from mobile operators [11], [12], vehicle location systems [13], [14] or population surveys [15] provide geolocated data containing some categorical variables. In all these cases, data vectors contain at least two continuous variables determining the latitude and longitude, and some other categorical variables which may take values belonging to a (non-necessarily ordered) finite set of labels.

In this paper we address the analysis of the relationship between non-ordered categorical variables and the geolocation ones. We begin by formalizing the problem statement in Section 2, where different alternatives for testing independence are also presented. These alternatives are computationally evaluated on two examples in Section 3. Finally, conclusions are presented in Section 4.

2. Problem Definition. Tests for Independence

Let us consider a set of measurements (x_i, y_i), $i = 1, \dots, N$ sampled from random variables (X, Y), where $X \in \mathbb{R}^2$ represents geolocation and $Y \in C = \{c_1, \dots, c_K\}$ is a categorical variable representing some property or feature. Each sample (x_i, y_i) represents an individual satisfying property y_i at location x_i.

As mentioned in the introduction, traditional tests to check the independence between random variables are defined for either categorical variables [2], [3] or continuous random variables or vectors [4]–[9]. Alternatively, some ANOVA-type tests are also available for cases where the dependent variable is continuous and the independent ones are categorical.

The problem of testing independence between a dependent categorical variable and independent continuous variables has been addressed mainly for the two categories (i.e., binary) case. When considering dependent categorical variables with more than two categories, binning of the continuous independent variable could be performed to apply tests between categorical variables. This procedure loses much information specially if the independent variable X is a vector. A k-nearest neighbors based procedure for estimating the mutual information has been recently proposed for the scalar X case [10], but its extension to vector cases has not been assessed. Hence, no tests seem to be consolidated for the case of dependent categorical variables (with more than two categories) and continuous independent vector variables, which happens to be the case under consideration.

In the following we begin by presenting some known ad hoc indexes which estimate spatial autocorrelation for labeled data in $X \in \mathbb{R}^2$.

2.1. Case Y Real Variable. Spatial Autocorrelation

For the cases that $Y \in \mathbb{R}$ (i.e., it represents a real value), spatial autocorrelation was formally addressed in [16] based on the work in [17], [18]. Since then, some improvements have been proposed [19]. If we denote $w_{i,j} = \frac{1}{d(x_i, x_j)}$ for $i \neq j$ and $w_{i,j} = 0, \forall i$, among the several measurements of spatial correlation that can be defined, the most common ones are *Moran's I* [17]:

$$I = \frac{N}{W} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{i,j} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{N} (y_i - \bar{y})^2}, \qquad \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i, \qquad W = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{i,j} \tag{1}$$

and *Geary's C* [18]:

$$C = \frac{N-1}{2W} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{i,j} (y_i - y_j)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \tag{2}$$

These indexes quantify space autocorrelation; hence, a test for the existence of such correlation can be formalized by estimating the corresponding p-values via a Bootstrap simulation procedure.

2.2. Categorical Variables

When $Y \in C = \{c_1, \dots, c_K\}$ only takes categorical values, the previous indexes cannot be directly defined unless an arbitrary number is assigned to label each class. Alternatively, the continuous variable $X \in \mathbb{R}^2$ can be binned so that standard independence test between categorical variables can be applied. In order to illustrate these procedures, a simple Mutual Information (MI) based index is proposed in the following section.

2.3. Mutual Information Based Index

The simplest procedure to estimate the mutual information index between X and Y relies on a discretization or binning of the continuous vector $X \in \mathbb{R}^2$. Let us consider a partition of the region of interest into a list of subsets $A_l \subset \mathbb{R}^2, l \in L$, so that each subset is assigned a label category $l \in L$. Hence, each location pair (x_{1i}, x_{2i}) is assigned the label l of the unique set A_l where it belongs to.

Then, the mutual information between the discretized and categorical variables can be computed as:

$$\mathrm{M{=}I(X,Y)}{=}\sum\nolimits_{l\in L}\sum\nolimits_{\mathrm{c_k}\in \mathrm{C'}} p(l,c_k)\log\left(\frac{p(l,c_k)}{p(l)p(c_k)}\right), \tag{3}$$

Note that this quantity can be normalized by the joint entropy $H(X,Y)$, but such normalization will not be necessary when computing the corresponding p-values associated with each case (via the Bootstrap procedure).

Obviously, this index will be sensitive to the discretization or binning of variable $X \in \mathbb{R}^2$. Therefore, new alternative statistics which avoid such approximations may valuable for testing independence. In the following, we present a statistic proposed in [20].

2.4. Herrera's Index

This index employs the information gathered in the categories or classes $c_k \in C' \subset C$ which have more than one element, i.e., such that $N_k = \#\{i \in \{i, \ldots, N\}: y_i = c_k\} > 1$. Defining an ordering among the measurements in each class $c_k \in C'$ (e.g., the one induced from the ordering of the whole set of measurements), and denoting k_i the (absolute) index in the whole set for the i-th element of class k, Herrera's index computes:

$$\mathrm{D}{=}\frac{\sum_{\mathrm{c_k}\in \mathrm{C'}}\sum_{\mathrm{i=2}}^{\mathrm{N_k}}\sum_{\mathrm{j=1}}^{\mathrm{i\text{-}1}}\mathrm{d}(\mathrm{x}_{k_i},\mathrm{x}_{k_j})}{\sum_{\mathrm{c_k}\in \mathrm{C'}}\sum_{\mathrm{i=2}}^{\mathrm{N_k}}(\mathrm{i\text{-}1})}{=}\frac{\sum_{\mathrm{c_k}\in \mathrm{C'}}\sum_{\mathrm{i=2}}^{\mathrm{N_k}}\sum_{\mathrm{j=1}}^{\mathrm{i\text{-}1}}\mathrm{d}(\mathrm{x}_{k_i},\mathrm{x}_{k_j})}{\sum_{\mathrm{c_k}\in \mathrm{C'}}\binom{\mathrm{N_k}}{2}} \tag{4}$$

In [20], the D index was computed for a random distribution of labels (preserving the N_k values) on the same x_i location values, and the ratio between this index D_r and the one obtained in (3) was provided as a final index: if the ratio is clearly larger than 1, this suggested that X and Y were not independent. Again, this heuristic procedure will be formalized in this work by estimating the p-value corresponding to D via an appropriate Bootstrap simulation scheme.

2.5. New Voronoi Based Entropy Index

Keeping in mind that $x_i \in \mathbb{R}^2$ the new method proposed in this paper first computes the Voronoi tessellation associated with the set $\{x_1, \ldots, x_N\}$. Let us call cell V_i the region associated with x_i. We can define that two cells V_i and V_j are adjoining if they share a common face, where only faces within the region of interest may be considered. Hence, the Voronoi tessellation allows the definition of a graph G so that each vertex represents a cell V_i with label y_i, and two vertices are connected if their corresponding cells are adjoining. Note that if we select those nodes with a given label value $y_i = c_k$ the corresponding subgraph G_k can be also defined.

For each subgraph G_k we compute the connected components CC_k^l of the graph, so that each CC_k^l, $l = 1, \ldots, L_k$ is again a subgraph of G_k. Denoting $|CC_k^l|$ the number of nodes of each connected component of G_k, we have that $\sum_{l=1}^{L_k}|CC_k^l| = N_k$, the number of nodes of subgraph G_k.

This new proposed ratio also employs the information gathered in such classes $c_k \in C'' \subset C$ having at least two elements, i.e., such that $N_k > 1$. It computes for each class the ratio between the entropy of the distribution of the sizes of the corresponding connected components and the maximal entropy associated with the size of such class. Then, all such entropy ratios are added to provide this new index E:

$$E = -\sum_{c_k \in C''} \frac{\sum_{l=1}^{L_k} \frac{1}{|CC_k^l|} \log(|CC_k^l|)}{\log(N_k)} \tag{5}$$

Once again, this new procedure can be also formalized by estimating the p-value corresponding to E via an appropriate Bootstrap simulation scheme.

3. Simulation Results

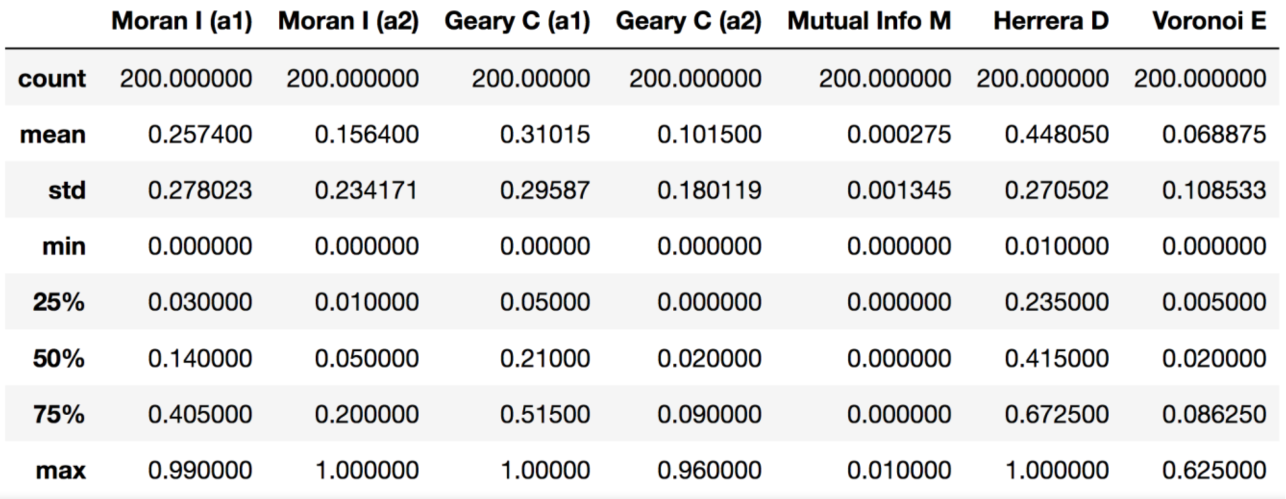
The following two examples illustrate different scenarios in order to assess the behavior of the different indexes presented.

3.1. Example 1

The indexes presented in equations (1)-(5) perform quite well in regions that cover similar ranges of distances in all directions. Nevertheless, the performance of some of these indexes deteriorates in regions which do not satisfy such regularity condition, as shown in the following example. To illustrate this sensitivity to the region characteristics, a high aspect ratio rectangular region with three ribbons each corresponding to a different class has been considered:

$$x = (x_1, x_2) \in S = [0,3] \times [0,100], \qquad f(x_1, x_2) = \begin{cases} A, if\ 0 \le x_1 \le 1, \\ B, if\ 1 \le x_1 \le 2, \\ C, if\ 2 \le x_1 \le 3. \end{cases} \tag{6}$$

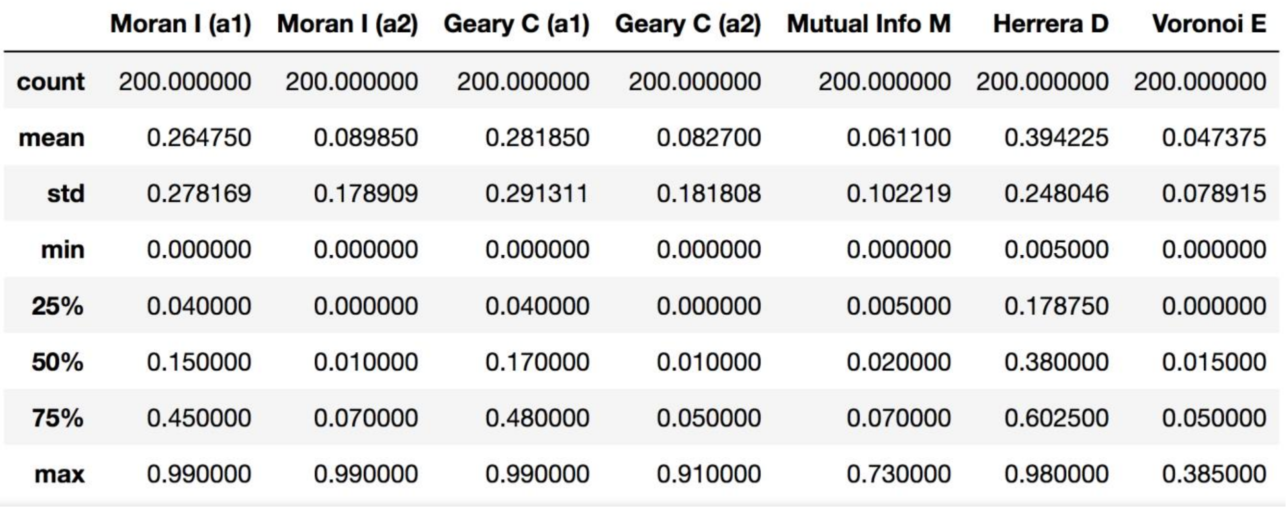
Two different number assignments to the labels, a_1 and a_2, will be considered so that: $a_1(A) = 0, a_1(B) = 1, a_1(C) = 2$ and $a_2(A) = 0, a_2(B) = 2, a_2(C) = 1$.



	Moran I (a1)	Moran I (a2)	Geary C (a1)	Geary C (a2)	Mutual Info M	Herrera D	Voronoi E
count	200.000000	200.000000	200.00000	200.000000	200.000000	200.000000	200.000000
mean	0.257400	0.156400	0.31015	0.101500	0.000275	0.448050	0.068875
std	0.278023	0.234171	0.29587	0.180119	0.001345	0.270502	0.108533
min	0.000000	0.000000	0.00000	0.000000	0.000000	0.010000	0.000000
25%	0.030000	0.010000	0.05000	0.000000	0.000000	0.235000	0.005000
50%	0.140000	0.050000	0.21000	0.020000	0.000000	0.415000	0.020000
75%	0.405000	0.200000	0.51500	0.090000	0.000000	0.672500	0.086250
max	0.990000	1.000000	1.00000	0.960000	0.010000	1.000000	0.625000

Fig. 1. Estimation of p-value distributions corresponding to basic Moran's I (with different assignment values for labels, a1 and a2), Geary's C (also with different assignment values for labels, a1 and a2), Herrera's ratio D and the new proposed Voronoi based entropy ratio index E for the region in Example 1. Number of points=120; Bootstrap samples=200, Montecarlo simulations=200; prob. of noise $\mathrm{p_n} = 0.2$.

A bootstrap technique, randomly shuffling the labels over the different geolocations, was employed to estimate the p-value associated to each index (Moran's I, Geary's C, Mutual Information M, Herrera's D and the Voronoi-based new proposed E). In addition, the distribution of such p-values was estimated via Montecarlo simulations. Some noise level was also incorporated in the data so that points belonging to a given region where assigned the correct label with probability $1-p_n$ and a label corresponding to each of the two alternative regions with probability $p_n/2$. Figure 1 displays the distribution of p-values provided by all the indexes.



	Moran I (a1)	Moran I (a2)	Geary C (a1)	Geary C (a2)	Mutual Info M	Herrera D	Voronoi E
count	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000	200.000000
mean	0.264750	0.089850	0.281850	0.082700	0.061100	0.394225	0.047375
std	0.278169	0.178909	0.291311	0.181808	0.102219	0.248046	0.078915
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.005000	0.000000
25%	0.040000	0.000000	0.040000	0.000000	0.005000	0.178750	0.000000
50%	0.150000	0.010000	0.170000	0.010000	0.020000	0.380000	0.015000
75%	0.450000	0.070000	0.480000	0.050000	0.070000	0.602500	0.050000
max	0.990000	0.990000	0.990000	0.910000	0.730000	0.980000	0.385000

Fig. 2. Estimation of p-value distributions corresponding to basic Moran's I (with different assignment values for labels, a1 and a2), Geary's C (also with different assignment values for labels, a1 and a2), Herrera's ratio D and the new proposed Voronoi based entropy ratio index E for the region in Example 2. Number of points=120; Bootstrap samples=200, Montecarlo simulations=200.

Simulations show that Moran and Geary's are sensitive to the numbers assigned a priori to each class (as well as to the region aspect ratio), whereas the Mutual Information based method performs very well. The behavior of Herrera's index D is quite poor, below the performance of Moran and Geary's indexes. Finally, the Voronoi based index E performs better than Moran and Geary's indexes, independently of the number assignment for these.

3.2. Example 2

This example illustrates the behavior of the indexes when the categorical labels are conditioned by some geolocated topological patterns such as streets or rivers:

$$x=(x_1,x_2)\in s=[-5,5]\times[-1,1], \qquad f(x_1,x_2)=\begin{cases} A, if\ |x_2-\frac{2}{5}x_1(\frac{x_1^2}{25}-1)|\leq 0.2, \\ B, \qquad if\ |x_1-x_2|\leq 0.2, \\ C, \qquad if\ (x_1,x_2)\ elsewhere. \end{cases} \tag{7}$$

Again, two different number assignments to the labels, a_1 and a_2, will be considered so that: $a_1(A)=0, a_1(B)=1, a_1(C)=2$ and $a_2(A)=0, a_2(B)=2, a_2(C)=1$.

The same bootstrap technique, randomly shuffling the labels over the different data geolocations, was employed to estimate the p-value associated to each index (Moran's I, Geary's C, Mutual Information M, Herrera's D and the Voronoi-based new proposed E). In addition, the distribution of such p-values was estimated via Montecarlo simulations. Figure 2 displays the distribution of p-values provided by all the indexes. Note that, in this case, the Voronoi-based new proposed method outperforms all the other procedures, since it is the most robust (smallest standard deviation) and most efficient (smallest average p-value) index. The good behavior of the Voronoi-based in this example may be due to the fact that the existing relationship between the variables is grounded on topological patterns which may not be detected so easily by distance-based schemes.

4. Concluding Remarks

The results for the new index based on the Voronoi tessellation for measuring the relationship between the non-ordered categorical variable and the geolocation variables are quite promising in terms of robustness and efficiency, deserving further research for a more complete characterization.

Funding

This work has been partially supported by project MTM2015-67396-P of Ministerio de Economía y Competitividad, Spain.

Acknowledgment

The authors want to thank Prof. to be J.F. Sánchez-Rada for his wise suggestions on the applicability of mutual information methods, and Dr. to be L. Úbeda-Medina for fruitful discussions on the subject.

References

[1] Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, *26(1)*, 97–107.

[2] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50(302)*, 157–175.

[3] McDonald, J. H. (2009). *Handbook of Biological Statistics*. Vol 2. Sparky House Publishing, Baltimore.

[4] Gretton, A., & Györfi, L. (2010). Consistent nonparametric tests of independence. *Journal of Machine Learning Research*, *11(Apr)*, 1391–1423.

[5] Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E, 69(6)*, 066138.

[6] Sricharan, K. Raich, R., & Hero, A. O. (2012). Estimation of nonlinear functionals of densities with confidence. *IEEE Transactions on Information Theory, 58(7)*, 4135–4159.

[7] Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with Hibert-Schmidt norms. *Proceedings of International Conference on Algorithmic Learning Theory* (pp. 63–77). Springer.

[8] Székely, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 2769–2794.

[9] García, J. E., & González-López, V. A. (2014). Independence tests for continuous random variables based on the longest increasing subsequence. *Journal of Multivariate Analysis*, *127*, 126–146.

[10] Ross, B. C. (2014). Mutual information between discrete and continuous data sets. *Plos One*, *9(2)*, e87357.

[11] Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, M., & Barabási, A-L. (2008), Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: mathematical and theoretical, 41(22)*, 224015.

[12] Zufiria, P. J., Pastor-Escuredo, D., Úbeda-Medina, L., Hernández-Medina, M. A., Barriales-Valbuena, I., Morales, A. J., Jacques, D. C., Nkwambi, W., Diop, M. B., Quinn, J., Hidalgo-Sanchís, P., & Luengo-Oroz., M. (2018). Identifying seasonal mobility profiles from anonymized and aggregated mobile phone data. Application in food security. *PLOS ONE*, *13(4)*, 1–20.

[13] Beraldi, P., & Bruni, M. E. (2009). A probabilistic model applied to emergency service vehicle location. *European Journal of Operational Research, 196(1)*, 323–331.

[14] Álvaro-Hermana, R., Fraile-Ardanuy, J., Zufiria, P. J., Luk Knapen, L., & Janssens, D. (2016). Peer to peer energy trading with electric vehicles. *IEEE Intelligent Transportation Systems Magazine, 8(3)*, 33–44.

[15] Chen, P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies. *Information Sciences, 275*, 314–347*.*

[16] Cliff, A. D., & Ord, K. (1970). Spatial autocorrelation: A review of existing and new measures and applications. *Economic Geography, 46(sup1)*, 269-292*.*

[17] Moran, P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, *10(2)*, 243–251.

[18] Geary, R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician, 5(3)*, 115-146*.*

[19] Burridge, P. (1980). On the cliff-ord test for spatial correlation. *Journal of the Royal Statistical Society, Series B (Methodological)*, 107–108.

[20] Herrera, C. (2017). Socio geographical patterns inferred from mobile phone records*.* PhD thesis, Universidad Politécnica de Madrid.

Pedro J. Zufiria and received the telecom. eng. degree by the Universidad Politécnica de Madrid (UPM) in 1986, the M.Sc. in ME, M.Sc. in EE, and Ph.D. degrees from the University of Southern California in 1989. He also received the doctor ingeniero de telecomunicación degree by the MEC in 1991 and the licenciado en ciencias matemáticas degree by the Universidad Complutense de Madrid in 1997. Since 1987 until 1990 he was teaching and research assistant in USC, and since 1991 he is professor in the UPM. His research interests focus on the analysis, control and fault diagnosis of dynamical systems, the theory of complex networks, and the study of learning paradigms for applications in data processing, having authored over 100 international publications in these fields.

Miguel Á. Hernández-Medina was born in Getafe, Spain, in 1964. He received a degree in mathematics from the Universidad Complutense de Madrid in 1987. In 1997 he received the Ph.D. degree in mathematics from the Universidad Politécnica de Madrid (UPM). Since 2000 he has been a professor at the E.T.S.I.T. of the Universidad Politécnica de Madrid. His research interests focus in mathematical sampling theory, harmonic analysis and the theory of complex networks.