STATISTICAL INFERENCE WITH SYNTHETIC FUNCTIONAL DATA DERIVED FROM POINT PROCESSES WITH APPLICATIONS

W. GONZÁLEZ-MANTEIGA, M. I. BORRAJO, AND I. FUENTES-SANTOS

Abstract. A common question when a given point process is observed in more than one population is whether those patterns share the same structure or they can be partitioned in a certain number of groups. To address this issue, recent advances on nonparametric inference for point processes are needed. In this talk we focus on kernel estimators of the first-order intensity and nonparametric tests for comparison of two point patterns.

Moreover, clustering algorithms, such as the k-means, can be used to classify a number of observed point patterns into groups. To tackle this problem we move from the point process framework with intensity functions, to the space of density functions. We describe the particularities of this space, and analyze the requirements, implementation and limitations of the k-means algorithm for classification of density functions.

The methodology presented is applied to different real data problems: COVID-19 infections and deaths in Spain, wildfires in Galicia (north-west Spain) and crime events in Rio de Janeiro (Brazil).

1. INTRODUCTION AND MOTIVATION

Point processes are mathematical models generating a random number of events on a measure space, $\mathcal{S}$. These processes appear in many different real problems in a wide variety of fields: ecology, geology, forestry, epidemiology, urban security,... A common problem in those scenarios is that of population comparison, i.e., several point patterns are observed in a same domain, and we are interested in determine whether those patters come from one, two or even more different populations. Formally, we want to group these patterns according to their underlying point processes.

To address these problem we will focus on three different illustrations in different fields:

- **COVID-19 infections and deaths:** in this context we observe unidimensional point patterns recording the number of cases on a certain period of time in different regions (Spanish provinces). We want to group the underlying processes generating those patterns in the different provinces according to their intensity, which in this case depends on a temporal variable and the synthetic data is then unidimensional.
- **wildfires:** we were provided with a very complete data set recording wildfires in Galicia (north-west Spain) during several years. In this data set, the cause of the wildfire was recorded, and we want to determine if the spatial distribution of wildfires in Galicia varies (or not) depending on this cause.

This work has been partially supported by Grant PID2020-116587GB-I00 funded by MCIN/AEI/ 10.13039/501100011033.

The talk at the 8IMM 2022 has been given by the first author.

- **crimes or road accidents:** these are actually two different examples, which share a common complexity, i.e., the observation domain. Both, road accidents and crimes occur on the road network, so, they 'do not live' any more on the euclidean space but on what is called a graph. This fact increases the complexity of the problem in both methodological and computational sides, but the idea is the same: we want to determine for example if the road accidents on a Monday are the same as on a Saturday, or if the spatial distribution of crimes with fatalities is the same of those where nobody was injured.

The main difference addressing this problem and a "classical" population comparison, is that in the latter the essential object for the comparison are the samples themselves, where in this new context, we are classifying intensity functions, more specifically estimations of intensity functions obtained from point patterns. Hence, we find ourselves working with synthetic data, particularly functional synthetic data.

As it has been defined by the *European data protection supervisor*, the concept of synthetic data generation is to take an original data source and create new, artificial data, with similar statistical properties. Keeping the statistical properties is crucial, and it means that by analysing the synthetic data, the same conclusions as those of analyzing the original data set should be drawn. The main idea behind synthetic data is that they are artificially created data rather than being generated by actual events, so they are not obtained by direct measurement. This is exactly what happens with our intensity estimates. The estimates do not come from direct observation, but from applying a certain procedure (estimation in this case) to the observations themselves.

2. NONPARAMETRIC METHODS FOR POINT PROCESSES

Defining mathematical tools to address this particular problem, requires a deep knowledge of point processes and nonparametric inference. Let X be a point process defined in a bounded region $W \subset \mathcal{S}$. Let $X_1, \ldots, X_N$ be a realisation of the point process where N is the random variable counting the number of events. The first-order intensity is defined as:

$$\lambda(x) = \lim_{|dx| \to 0} \frac{\mathbb{E}[N(dx)]}{|dx|},$$

where $|dx|$ denotes the measure of an infinitesimal region containing the point $x \in W \subset \mathcal{S}$. Intuitively, $\lambda(x)$ measures the expected number of events per unit measure and, consequently, characterizes the distribution of the point process. This is a non-negative function whose integral is the expected number of events, so, relying on this idea, it is almost straightforward to build a sort of "artificial density function", known as the density of events locations:

$$\lambda_0(x) = \lambda(x)/m,$$

where the intensity factor, $m = \int_W \lambda(x)dx$, is the expected number of events lying on W.

Both the intensity and density of event locations can be estimated by kernel smoothing. In the euclidean space, the kernel estimator of the density of event locations is

$$\hat{\lambda}_{0,H}(x) = (p_H(x)N)^{-1} |H|^{-1/2} \sum_{i=1}^{N} K\left(H^{-1/2}(x - X_i)\right) 1[N \neq 0] \tag{1}$$

where $1[\cdot]$ is the indicator function, H is a bandwidth parameter, K denotes a kernel function, $K_H(x) = |H|^{-1/2}K\left(H^{-1/2}x\right)$, and $p_H = \int_W |H|^{-1/2}K(H^{-1/2}(x-y))dy$ is an edge correction term. When $d > 1$, the bandwidth parameter, H, is a symmetric and positive-definite matrix and $|H|$ is the determinant of H. [2] proved the consistency of (1) and proposed a plug-in bandwidth selector for inhomogeneous spatial point processes.

Taking advantage of the consistency of $\hat{\lambda}_{0,H}(x)$, and of the relationship between the density of event locations and the density of multivariate distributions in $\mathbb{R}^d$, inference tools developed for the latter can be extended to the point process framework. In this line, a Cramer-von Mises and a no effect test have been proposed to compare the distribution of two spatial point processes, see details in [3]. Moreover, [1] addressed the particular scenario of the two-sample problem for point processes with covariates.

3. CLASSIFICATION ALGORITHM FOR POINT PROCESSES

Classification is the problem of identifying which of a set of categories or groups, an observation (or observations) belongs to. Focusing on point processes observed in the one-dimensional Euclidean space for simplicity, our aim is to classify a set of Poisson point processes $\{X_i\}_{i=1}^n$ observed in a given interval $S = [a, b] \subset \mathbb{R}$ into K groups, conditioning on point processes in each group share a same distribution. To address this problem, we estimate the corresponding first-order intensity function $\{\hat{\lambda}_i(x)\}_{i=1}^n$, moving from the point process framework to the intensity space, Ω. As argued by [4], this space can be seen as a product metric space $\Omega = \mathcal{D} \times \Omega_S$, where $\mathcal{D} \subset \{f : S \to \mathbb{R}^+; \int_S f(x)dx = 1\}$ denotes the spaces of density functions in S, and $\Omega_S = \mathbb{R}^+$ the space of intensity factors, which determine the shape and expected number of events (size) of the point process, respectively.

Therefore, we can use a L^2 product metric, d, between a given pair of intensity functions $\lambda_1 = (m_1, f_2)$ and $\lambda_2 = (m_2, f_2)$ given by

$$d(\lambda_1, \lambda_2) = \left(d_{\mathcal{D}}^2(f_1, f_2) + d_E^2(m_1, m_2)\right)^{1/2}, \tag{2}$$

where d_E is the one-dimensional Euclidean metric and $d_{\mathcal{D}}$ is a metric in the density space. Considering this decomposition, the structure of point processes relies on the density of event locations and, consequently, our problem can be reduced to that of classifying density estimates in groups.

Let $\{X_i\}_{i=1}^n$ be a set of point patterns observed in a bonded interval S, and $\{\hat{f}_i(x), x \in S\}_{i=1}^n$ the kernel estimators of their densities of event locations. Let assume that these point processes belong to K categories characterized by the densities of events locations $\{f_k\}_{k=1}^K$, referred as centers. Classification of the n density estimates into the K groups can be conducted by a k-means algorithm in the space of density functions, $\mathcal{D}$. Intuitively, the density estimates can be considered as functional data, and the k-means algorithm for functional data could be used to proceed with classification. However, $\mathcal{D}$ is not a Hilbert space, and consequently, statistical procedures for functional data evaluated on a Hilbert space can not be directly applied here. In particular, we cannot use the L^2 distance as discrepancy measure in the k-means algorithm.

Following a common practice for statistical modeling and computing of densities, we should conduct the classification in a representative space. These spaces have been mainly defined under two perspectives, the functional and object-oriented approaches, see details in [5]. In this work we use both of them; we propose a transformation approach for functional data representation, as well as two object-oriented metrics to determine the discrepancy between density functions. The selection of these metrics is also a crutial point in the development of this methodology.

- **Transformation approach (L^2-LQD)**: density curves can be treated as functional data after transformation into a Hilbert space. Here we use the log-quantile density (LQD) transformation, and the L^2-distance in the transformed space:

$$d_{LQD}(f_1, f_2) = \left(\int_0^1 \left(\psi_{LQD(f_1)}(x) - \psi_{LQD(f_2)}(x) \right)^2 dx \right)^{1/2},$$

  where $\psi_{LQD(f)}(\cdot)$ denotes the LQD-transformed density.
- **L^2-Wasserstein distance (L^2-WS)**: this is an optimal transport distance that measures the cost of transporting one distribution to another in the object-oriented framework that can be defined in general spaces. For absolutely continuous distributions, it can be defined as the L^2-distance between the corresponding quantile functions, $Q_{f_j}, j = 1, 2$:

$$d_W(f_1, f_2) = \left(\int_0^1 \left(Q_{f_1}(r) - Q_{f_2}(r) \right)^2 dr \right)^{1/2}.$$

- **Fisher-Rao distance (FR)**: this is the spherical geodesic distance between square root densities:

$$d_{FR}(f_1, f_2) = \arccos \left(\int_a^b \sqrt{f_1(x) f_2(x)} dx) \right),$$

  where arccos denotes the arco-cosine function. The square root of a density lies on the Hilbert unit sphere, so d_{FR} measures the length of an arch connecting $\sqrt{f_1}$ and $\sqrt{f_2}$ along this sphere.

In addition to the selection of the metric, the determination of the number of groups is also crucial. Assuming that K is known is not realistic in practice, and can lead to a misclassification. For this reason, calibration procedures to determine the number of groups prior to run the classification algorithm are required. This problem has already been addressed in [6] for regression curves and it is an ongoing work we are tackling for intensity estimates.

4. APPLICATION TO COVID-19

One of the main motivations for tackling this problem is the COVID-19 pandemic. Since March 2020 we have been suffering one of the most important sanitary emergencies of the last centuries. Spain was very affected by this pandemic, specially in its beginning. Our national health system is decentralised, i.e., every region manages its own health system, even though they are given directions from the national government.

An important administrative division in Spain are provinces (the level below regions), and we are interesting in comparing the behaviour of the COVID-19 infections and deaths

in these provinces. We have gather daily records of infections and deaths in every province from March 2020 until March 2022 (from this point on, the way of counting cases and deaths has changed in Spain). In Figure 1 we can see the total number of cases and deaths per province in Spain.

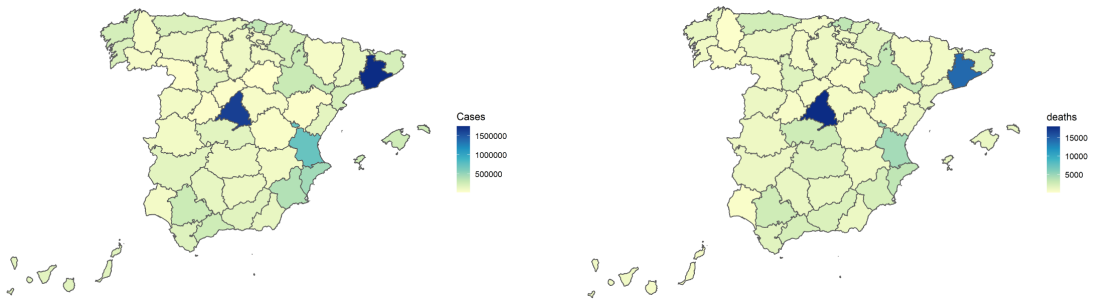

FIGURE 1. Total number of COVID-19 cases (left) and deaths (right) per province in Spain from March 2020 to March 2022.

For each province, with the temporal point patterns given by the daily number of cases (deaths) we have estimated the corresponding density of event location using kernel methods as detailed in Section 2. These estimated curves, for both cases and deaths, can be seen in Figure 2.

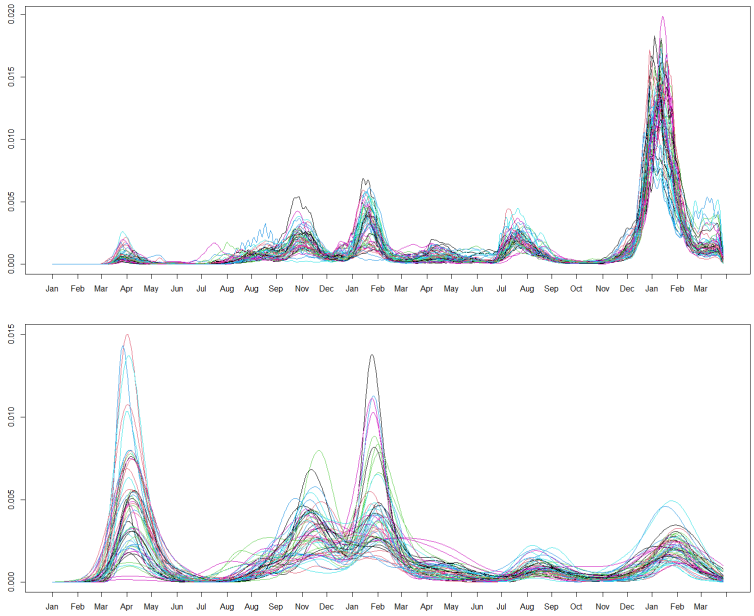

FIGURE 2. Temporal density estimates of cases (top) and deaths (bottom) per province in Spain from March 2020 to March 2022.

After applying the classification methods with the appropriate distance measures, the results of the grouped densities can be seen in Figure 3.

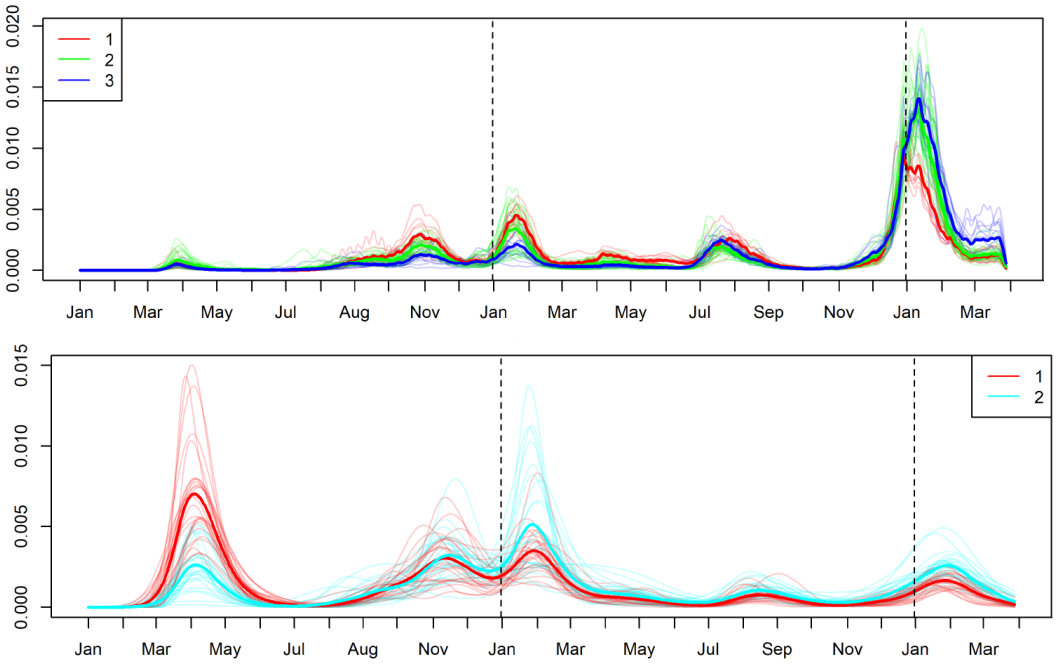

FIGURE 3. Result of the classification in groups of the temporal intensity functions of cases (top) and deaths (bottom).

These results can be easily translated into Spanish administrative map in order to see the spatial distribution of these groups. These results are presented in Figure 4.

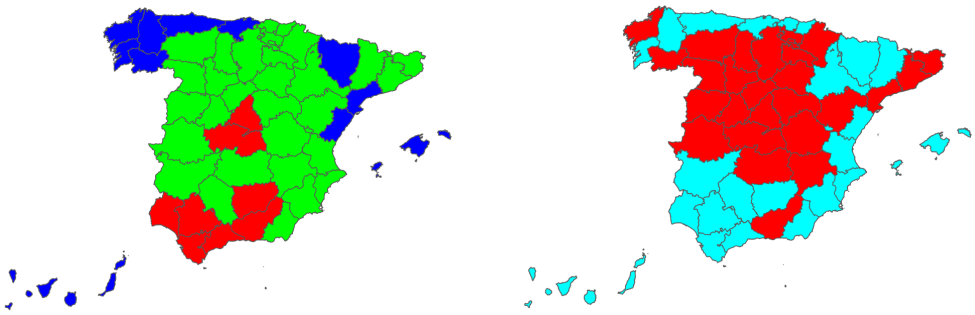

FIGURE 4. Result of the classification algorithm for cases (left) and deaths (right).

We have classified COVID-19 cases into 3 groups (Figure 3, top), provinces in group 1 (red) suffered higher incidence during the second and third wave than those in groups 2

(green) and 3 (blue), but lower incidence during the last wave. Provinces in group 3 reported the lowest incidence during the second and third wave and a slower incidence decrease than group 2 in the last wave.

Figure 3 (bottom), shows that the classification of deaths patterns is dominated by the first wave, where provinces in group 1 suffered higher incidence than those in group 2. Once classified the infection and death patterns and identified the provinces in each group (Figure 4), our next aim is to check the effect of demographic and socioeconomic factors, as well as the different COVID-19 mitigation strategies, on those patterns.

References

[1] Borrajo, M. I., González-Manteiga, W., and Martínez-Miranda, M. D. (2020). Testing for significant differences between two spatial patterns using covariates. *Spatial Statistics*, 40.

[2] Fuentes-Santos, I., González-Manteiga, W., and Mateu, (2016). Consistent smooth bootstrap kernel intensity estimation for inhomogeneous spatial poisson point processes. *Scandinavian Journal of Statistics*, 43(2), 416-435.

[3] Fuentes-Santos, I., González-Manteiga, W., and Mateu, J. (2021). Testing similarity between first order intensities of spatial point processes. A comparative study. *Communications in Statistics- Simulation and Computation*, 1-21.

[4] Gajardo, A., and Müller, H. G. (2021). Point process models for COVID-19 cases and deaths. *Journal of Applied Statistics*, 1-16.

[5] Petersen, A., Zhang, C., and Kokoszka, P. (2022). Modeling probability density functions as data objects. *Econometrics and Statistics*, 21, 159-178.

[6] Vogt, M., and Linton, O. (2017). Classification of non-parametric regression functions in longitudinal data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1), 5-27.

W. González-Manteiga; Universidade de Santiago de Compostela.
Email address: wenceslao.gonzalez@usc.es

M.I. Borrajo; Universidade de Santiago de Compostela.
Email address: mariaisabel.borrajo@usc.es

I. Fuentes-Santos; Instituto de Investigaciones Marinas (IIM-CSIC).
Email address: isafusa@iim.csic.es