Filtering and Mapping Public Health Data with an Innovative Kriging Approach, Accounting for Single Observation Variance

Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 26 (2015) 57 – 61

Procedia Environmental Sciences

Spatial Statistics 2015: Emerging Patterns

Filtering and mapping public health data with an innovative kriging approach, accounting for single observation variance

Vittorio Casella[a,*], Ambrogio Manzino[b], Riccardo Bellazzi[c], Marica Franzini[a]

[a]*University of Pavia – Department of Civil Engineering and Architecture, Pavia, Italy*
[b]*Polytechnic of Torino – Department of Environment, Land and Infrastructure Engineering, Torino, Italy*
[c]*University of Pavia – Department of Electrical, Computer and Biomedical Engineering, Pavia, Italy*

Abstract

The main scope of the paper is performing appropriate kriging interpolation of the diabetes prevalence data coming from the Pavia (Italy) Local Health Care Agency (ASL). The original dataset is analyzed, the Bayesian regularization is evaluated, which is applied by other authors and finally prevalence data are simulated by means of random fields, in order to tune and evaluate kriging interpolation.

Peer-review under responsibility of Spatial Statistics 2015: Emerging Patterns committee

Keywords: kriging; spatial epidemiology; diabetes.

1. Introduction

The paper is developed in the frame of ongoing exposomics (a novel discipline aiming to correlate health with the factors which people are exposed to: food, lifestyle and environment) activities at the University of Pavia, Italy. Spatial statistics analysis of public health variants is performed, in particular kriging interpolation. The Province of Pavia (3000 sq. km wide, 550000 inhabitants and constituted by 190 municipalities) is adopted as a pilot. The present paper deals with diabetes prevalence data and illustrates the dataset, in Sec. 2; evaluates the Bayesian regularization which is often adopted in similar studies, in Sec. 3; simulates the prevalence data by means of random fields in order to tune and evaluate a peculiar kriging interpolation, in Sec. 4.

*Corresponding author. Tel.: +39 0382 985417; fax: +39 0382 985419.
E-mail address: vittorio.casella@unipv.it

Peer-review under responsibility of Spatial Statistics 2015: Emerging Patterns Committee
doi:10.1016/j.proenv.2015.05.024

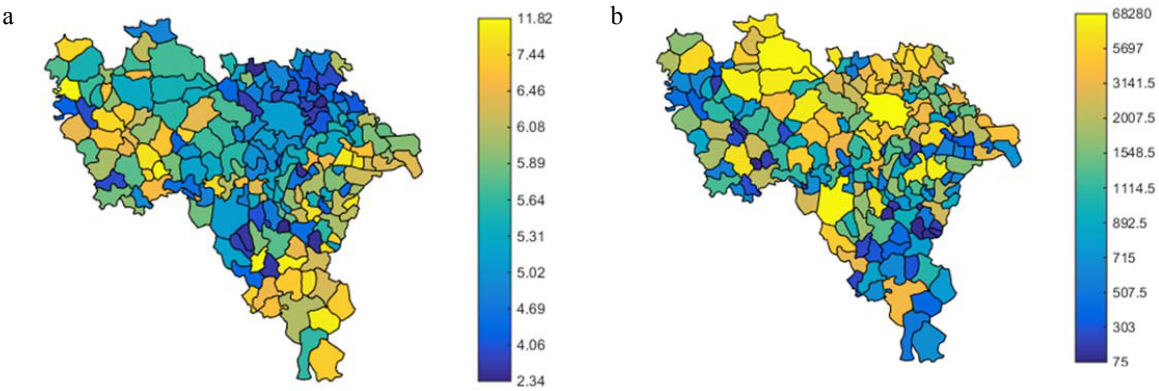


Fig. 1. (a) Geographic distribution of diabetes prevalence over the 190 municipalities, color scale in based on percentiles; (b) number of inhabitants for each municipality.

a

Descriptive statistics figures for prevalence data (F) supplied by the Local Health Care Agency	
Number of municipalities	190
Min [%]	2.34
Max [%]	11.82
Range [%]	9.48
Mean [%]	5.73
STD [%]	1.44
Median [%]	5.64
Average mean (weight is the number n of inhabitants) [%]	5.39

b

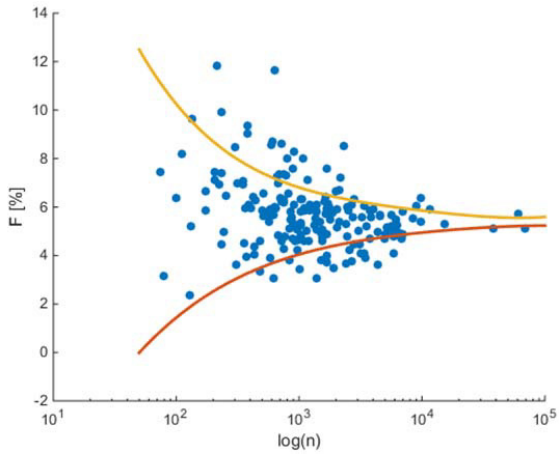


Fig. 2. (a) Descriptive statistics for the considered variant; (b) funnel plot of the 95% confidence interval for the *F* estimator as a function of population (n); blue dots represent the 190 municipalities.

2. The data analyzed and their stochastic properties

The Local Health Care Agency (ASL) yearly collects approximately 100 public health data variants for each of the 190 municipalities composing the Province of Pavia. Following analysis concerns the diabetes prevalence for the 2010 year. By *prevalence,* we mean the number of sick people of a certain territory and a defined time span, as a fraction of the total number of the inhabitants; it can be measured in absolute terms or percentage units, as we will do. Geographic distribution of diabetes prevalence is shown in Fig. 1a while Fig. 1b illustrates the number of inhabitants, highlighting that several municipalities are poorly populated: 2 have population lower than 100, and 37 have less than 500.

Descriptive statistics is shown in Fig. 2a, highlighting that the mean values is significantly lower than mid-range and therefore the histogram is skewed.

Following the existing literature (Martuzzi and Elliot [1]; Berke [2]), we assume that the number of sick people (the variant originally observed) in each municipality follows the binomial distribution $P_B(k;n,p)$; prevalence figures come from a statistical yet simple estimator F whose uncertainty depends on the number of inhabitants n : $E[F]=p$, $VAR[F]=p(1-p)/n$. As a consequence, the variability shown in Fig. 1a is the *sum* of a spatially correlated signal (in some regions spatial correlation is clearly visible) with an uncorrelated random noise whose size depends on population. The funnel plot Fig. 2b is calculated assuming a constant true prevalence of 5.4% for all

the Province and shows the lower and upper bounds of the 95% confidence interval. The plot shows that much of the

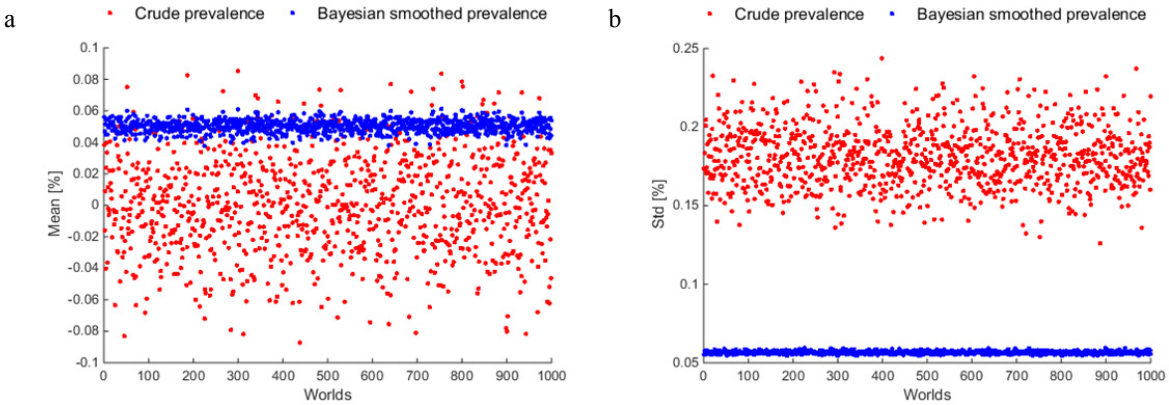
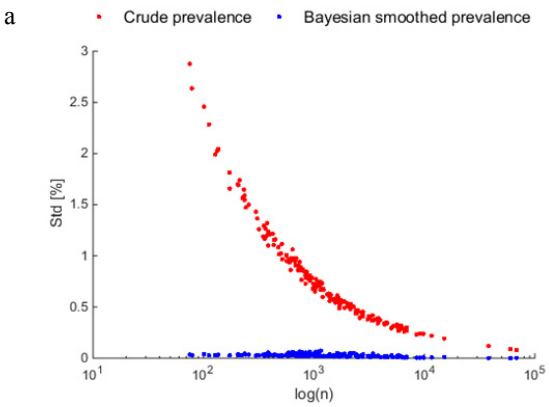


Fig. 3. Column wise analysis. (a) Mean values of the differences between estimated and true prevalence, for each world simulated; (b) standard deviation of the differences.



b

	Min	Max	Range
True prevalence [%]	4.46	7.54	3.09
Estimated crude prevalence [%]	4.47	7.60	3.14
Bayes smoothed prevalence [%]	4.30	6.44	2.14

Fig. 4. Row-wise analysis. (a) Standard deviation of the estimations, as a function of the population logarithm; (b) min, max and range of the differences between row wise averaged prevalence estimations and true values.

prevalence variability is due to the estimator behavior; nevertheless, as many points are outside the confidence interval, it also suggests that the true prevalence is variable over the considered territory.

3. The Bayesian regularization and its effects

Because of the significant variance variability shown by the so-called crude prevalence, illustrated in Fig. 2b, several authors (Berke [2]; Marshall [3]) apply Bayesian smoothing before further analyzing data. We adopted the methodology proposed by (Martuzzi and Elliot [1]) and implemented it in a Matlab function; the so-coded estimator will be indicated with F^B .

We simulated diabetes prevalence over the Province of Pavia by means of suitably-created Matlab functionalities. First of all, the crude prevalence data coming from the Local Health Care Agency were georeferenced and interpolated with a smoothing methodology, in order to create a likely continuous surface of *true* prevalence; then each municipality was associated with its true prevalence and population; successively a column vector made of n_a (the number of municipalities) values containing the simulated number of sick people was calculated; the procedure was iterated n_w times in order to create a $n_a \times n_w$ matrix representing n_w simulated worlds. Furthermore, the F

and F^B estimators were applied to all columns of the generated matrix, obtaining two further matrices of the same size containing the estimated crude and smoothed prevalence.

The so-obtained prevalence matrices were in-depth analyzed, by performing column wise and row wise analysis. Fig. 3 shows one example of the first type: in abscissa there are the 1000 worlds considered; in ordinate there are, in Fig. 3a; the mean value of the differences between estimated and true prevalence, represented by red dots for crude prevalence and by blue ones for Bayes-smoothed figures; Fig. 3b shows standard deviation of the differences.

The Bayes smoother is slightly biased (Fig. 3a), less noisy on average (Fig. 3b) and its variance is less dependent from the population (Fig. 4a). Nevertheless we came to the conclusion that it loses information: we averaged row wise thus coming to a reliable estimation of prevalence for each municipality; Fig. 4b reports minimum and maximum values, together with their range, for the true (simulated) prevalence and for its crude and Bayes smoothed estimations. The latter shows a 30% decrease of the range and this highlights, in our opinion, an information loss.

4. Random fields simulation and analysis

Kriging interpolation of prevalence data must be conceptually framed as follows: prevalence is represented by a second-order stationary, self-correlated surface; if we had punctual samplings from that surface, we could apply kriging directly to them. The Local Health Care Agency collects different samplings, corresponding to the average values of the above cited surface calculated upon the territory of the various municipalities. Moreover, those samplings are obtained through an estimator whose variance significantly depends on the population. In summary, prevalence data coming from the Agency can't be directly interpolated by kriging.

In order to support its development, a Matlab-based simulation tool was developed, which is able to generate random fields with user defined mean, variance and auto-covariance function. Primarily, the random field representing prevalence is discretized, thus forming a matrix random variable (RV); the elements of that matrix are mapped to a column random variable; variance-covariance matrix of the latter is calculated taking into account the user defined auto-covariance function and the relative distances of its elements (distances are evaluated going back to the matrix RV). Finally an extraction from that column random variable with given mean and covariance matrix is obtained by the suitable Matlab function and projected back to the matrix form, as Fig. 5a shows. Noticeably, the procedure can be used to generate the averaged samplings and to simulate estimation errors. Thus, a likely model of the data coming from the Local Health Care Agency was created and will be used to define, tune and evaluate suitable kriging procedures. This will be the topic of a further paper.

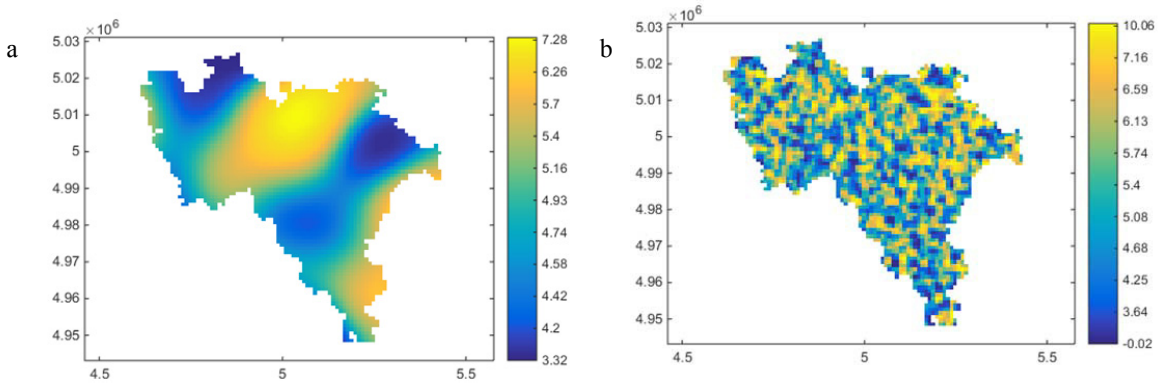


Fig. 5. (a) Simulation of the discretized random field with $\mu = 5.4, \sigma = 1.4$ and $\sigma_{\text{corr}} = 15\,\text{km}$; (b) simulation with $\sigma_{\text{corr}} = 1\,\text{km}$: as auto-covariance diminishes, then simulated RV resembles white noise.

References

1. Martuzzi M, Elliot P. Empirical Bayes estimation of small area prevalence al non-rare conditions. *Statistics in Medicine* 1996; **15**: 1867-1873.
2. Berke O. Exploratory disease mapping: kriging the spatial risk function from regional count data. *International Journal of Health Geographics* 2004; **3**.
3. Marshall RJ. Mapping disease and mortality rates using empirical Bayes estimators. *Applied Statistics* 1991; **40**: 283-294.