

UNIVERSITÀ COMMERCIALE "LUIGI BOCCONI" – MILANO

Facoltà di Economia

Dottorato di Ricerca in Statistica

XVII Ciclo

Essays in Spatial Data Modeling

Coordinatore:

Ch.mo Prof. Pietro Muliere

Tesi di:

Massimiliano Copetti

n.matr. 901660DT

UNIVERSITÀ COMMERCIALE "LUIGI BOCCONI"
ISTITUTO DI METODI QUANTITATIVI

The thesis "**Essays in Spatial Data Modeling**"
by **Massimiliano Copetti** n.matr. 901660DT
is recommended for acceptance by the members of the
delegated committee, as stated by the enclosed reports,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy.

Dated: January 2007

Research Supervisor: **Francesco Billari**

External Examiners: **Peter J. Diggle**
Giuseppe Arbia

UNIVERSITÀ COMMERCIALE "LUIGI BOCCONI"

Date: **January 2007**

Author: **Massimiliano Copetti** **n.matr.**

901660DT

Title: **Essays in Spatial Data Modeling**

Department: **Istituto di Metodi Quantitativi**

Permission is herewith granted to Università Commerciale "Luigi Bocconi" to circulate and to have copied for non-commercial purposes, at its discretion, the above title upon the request of individuals or institutions.

Signature of Author

THE AUTHOR RESERVES OTHER PUBLICATION RIGHTS, AND NEITHER THE THESIS NOR EXTENSIVE EXTRACTS FROM IT MAY BE PRINTED OR OTHERWISE REPRODUCED WITHOUT THE AUTHOR'S WRITTEN PERMISSION.

THE AUTHOR ATTESTS THAT PERMISSION HAS BEEN OBTAINED FOR THE USE OF ANY COPYRIGHTED MATERIAL APPEARING IN THIS THESIS (OTHER THAN BRIEF EXCERPTS REQUIRING ONLY PROPER ACKNOWLEDGEMENT IN SCHOLARLY WRITING) AND THAT ALL SUCH USE IS CLEARLY ACKNOWLEDGED.

I am my father.

Table of Contents

Table of Contents	ix
Preface	xi
Acknowledgements	xiii
1 Statistical Analysis of Innovation's Geography in Italy	1
1.1 Introduction	1
1.2 Data	2
1.3 Point Processes	4
1.3.1 Point Process Theory	4
1.3.2 Poisson Point Processes	8
1.3.3 Cox Process	10
1.3.4 Remarks: Inhomogeneous Poisson Process and Cox Process	11
1.4 Classical Spatial Analysis	12
1.4.1 Some notes on how I chose the optimal bandwidth .	15
1.5 Spatial Segregation	20
1.5.1 Spatial Segregation Theory	20
1.5.2 Application and Tests proposed	22
1.5.3 Monte Carlo Global Test	28
1.5.4 Conclusions	28
1.6 Temporal Analysis	29
1.6.1 Monte Carlo Test: same distribution among sectors over time?	29
1.6.2 Temporal Segregation	30
1.6.3 Intensity Kernel Estimator	36
1.6.4 Conclutions	37

2	Semiparametric Intensity and Density Estimation	43
2.1	From Density to Intensity	43
2.2	Semiparametric Density Estimation	44
2.2.1	Introduction	44
2.2.2	Nonparametric correction on a fixed start	45
2.2.3	Nonparametric correction on a parametric start	46
2.2.4	Comparison with the traditional kernel density estimator	50
2.2.5	Choosing Smoothing Parameter	51
2.2.6	Accuracy of the estimated correction factor	53
2.3	Semiparametric Intensity Estimation	54
2.3.1	Introduction	54
2.3.2	A parametric function corrected by a kernel factor	54
2.3.3	Choosing the Optimal Bandwidth	56
2.3.4	Simulation Study	56
2.3.5	Conclusions	64
2.4	Semiparametric Intensities Estimation using <i>pooled</i> data	64
2.4.1	Introduction	64
2.4.2	<i>Pooled</i> data: the semiparametric approach	64
2.4.3	Simulation Study	66
3	Spatial dependence in the tails of NO2 distributions: an extreme value theory approach	73
3.1	Introduction	73
3.2	Spatial Joint Risk	74
3.3	Spatial Tail Dependogram	76
3.4	Nonparametric Spatial Dependogram Estimation	77
3.5	Parametric Inference for Dependogram: an extreme value theory approach	80
3.5.1	Univariate threshold model	80
3.5.2	Bivariate threshold model	82
3.5.3	Choice of the threshold	85
3.6	Empirical Application: the NO2 case in Rome	85
3.7	Conclusions	92
	Bibliography	95

Preface

The present thesis is composed of three main chapters which can be seen as independent works too. The factor that they have in common is the spatial nature of the data analyzed.

The first chapter deals about a statistical analysis of the innovation in economics in Italy. Patents' spatial distribution in Italy seems to show some Innovation's clusters. A deeper view needs to consider the multivariate nature of this phenomenon: each patent belongs to an industrial sector. Considering each of the six sectors individually, they show, more or less, the same clusters which are present in the global distribution too. So a kernel regression estimator (as in Diggle *et al.*, 2005, (17)) was applied, putting each sector relative to the others, to describe the *spatial segregation* allowing us to distinguish the proper sectorial districts. Tools for testing the hypothesis of no segregation and the statistical significance of the estimates are provided. The procedure was easily extended to the temporal dimension using the application's date allowing to detect proper increase or decrease in the specific sectors' innovation activity.

In the second chapter a semiparametric estimator for the intensity of the Poisson point process is proposed. It combines a parametric start with a non-parametric kernel correction factor. It's not necessarily that the initial parametric estimate provides a serious approximation to the true intensity. In this case this method works well as the totally nonparametric kernel one. It works better if the parametric initial estimate is closed to the true intensity. The value of the optimal bandwidth and the plot of the correction factor could suggest us if and where the parametric part doesn't fit well

allowing us to make a better parametric hypothesis.

The new estimator is very useful for *pooled* data, typically generated by a multivariate Poisson point process. In this setting the parametric part is estimated using the whole dataset, while the non-parametric correction factor just using the specific subdataset. Here the role of the correction is different: it states if and where the specific component process departs from the shared common pattern.

The third work shows the use of extreme values theory for exploratory spatial analysis. The study of the pollutants needs a better understanding of their extreme behaviours which could potentially cause adverse health effects. Analysing spatial dependence of the pollutant, the dependogram proposed by Arbia and Lafratta (2005,(1)) is to be preferred to the correlogram because it captures non-linear relations in the tails of the joint distributions to better detect a pattern of spatial regularities. An advanced inference framework is proposed here to estimate the spatial dependogram: the estimated conditional distribution function is obtained with the application of a threshold-model, since we are interested much more in a robust right-tail distribution estimation. The method is applied to the data collect by 7 monitoring sites in Rome during the year 2000 e 2001.

Acknowledgements

I'm very grateful with the 'L. Bocconi' University and the Institute of Quantitative Methods which gave me the opportunity to attend the PhD program in statistics. In particular, I owe to prof. Pietro Muliere and to prof. Francesco Billari: they never have stopped to give me give encouragements and suggestions.

I would like to thanks prof. Stefano Breschi for providing me patents' dataset.

With a great pleasure I would like to acknowledge prof. Peter J. Diggle for the wonderful period I spent in Lancaster working on the first two chapters of this thesis. I thank prof. Giuseppe Arbia who is supervising me since my undergraduate studies: the third chapter is part of the work I did with him during the last year having a post-doc position at 'G. d'Annunzio' University. The last year has seen increasing the deep friendship with dr. Giovanni Lafratta, bynow my older brother.

My PhD colleagues, become closed friend, are in this thesis, with their comments and stimulating discussion. In particular Raffaele, Sergio and Gabriella in Milan, Gerwin and Nick in Lancaster and Annalisa in Genoa have left a deep mark in the days spent to live together.

This work would not have been possible without the costant presence of my mother, the tower of strenght of my family.

The last four years, the PhD years, have been the worst ones of my personal life. But something is changing and now I'm happy for at least three reasons.

My thesis has been completed.

I have met in these years my *maestro*, professor Peter J. Diggle.

I have met Stefania and her unbounded love.

Chapter 1

Statistical Analysis of Innovation's Geography in Italy

1.1 Introduction

The recent resurgence of growth studies has indicated technological activities as one of the most important factors in determining the performance of the economic systems, and a greater attention has been devoted to the geographical dimension of the mechanisms of creation and diffusion of technology. According to economics' tradition (Jaffe et al., 1993 (20); Jaffe et al., 2000 (19); Thompson&Fox-Kean, 2002 (24); Breschi&Lissoni, 2004 (7)), we represent Innovation as patents' production and we take the inventor's residence as the patent's location. In such a way, we can interpret the spatial pattern of patents as the realization of a spatial point process - more precisely, of a multivariate spatial point process - because each patent, i.e. each point, is characterized by the industrial sector which it belongs to: so that we have as many component processes as the number of industrial sectors.

It could be useful to estimate and map the spatial distribution of the total point process, or the intensity of each component process, but, we think, it's more interesting to study the behaviour of each component process - each sector - relative to the others.

Therefore, methodological problems of intensity estimation arise from the multivariate nature of the patents point pattern and the answer to this problem can be given analyzing the *Spatial Segregation* of the phenomenon under study which becomes the object of our inference through the so-called *type-specific probabilities* $p_k(x)$ (Diggle et al., 2005 (17)), representing the conditional probability that a case occurring at location x were of type k , which are built starting from ratios of intensities of component processes. A kernel regression estimator is proposed to make inference about the type-specific probabilities as in Diggle et al., 2005 (17)

A cluster detected in a single component map could be not informative or not interesting if it is present in other component maps too. Instead, a peak detected in the type-specific probability map of one sector means a significant and real predominance of that sector on the others.

We are also able to give a temporal dimension to each patent, the application date, allowing us to extend the previous analysis to the temporal setting detecting, possibly, *temporal segregation*.

A statistical test against the null hypothesis of no segregation is also provided. both in the spatial and in the temporal cases.

The EPO (European Patent Office) dataset built as CESPRI - Bocconi University is used to analyse the case of North-Italy.

1.2 Data

The EPO-Dataset shows all patent applications made at the European Patent Office (born in 1978 by Monaco's Convention) from 1978 by firms, laboratories, universities

of all Countries to get intellectual property and protection to their inventions in Europe. We use its version built at CESPRI - Bocconi University - Milan. The dataset gives us information about petitioners, inventors, date of request (day-level), International Patent Classification (IPC) code to distinguish the industrial sector and citations among patents. Using the residence of inventors it was possible to assign a precise spatial location. In this paper we consider the case of Italy in nineties. We omitted to consider the eighties because this period has been a transitional and experimental one for the Italian inventors as regard this new type of patent.

We omitted data from Sicily, Sardinia and the other isles: in fact using Euclidean distance we could have caused problems when interpreting the results, so we consider only the Italian peninsula. The loss of data is negligible.

The resulting dataset display 44078 inventors but only 25312 patents: this shows the presence of multi-inventors patents. To avoid any artificial way to assign a unique location to that patents, we considered only patents with one inventor leading to 14632 observations. This choice will influence the interpretation of results and our conclusions.

The whole North is very productive and it captures more than 83% of all patents! The middle-italy captures the 14%, mostly around Rome and Florence. A small share (3%) is present in the South, almost all around Naples. These findings don't surprise us: the industrial gap among Italian regions is well-known. This consideration suggested us to restrict to study just the North of Italy where the phenomenon seems to be more structured.

Two kinds of temporal dates are available: the publication date and the priority date, i.e. the date of the earliest filling of an application in any of the patent offices adhering to the convention. The choice is crucial because the time lag between the

priority date and the publication date may range from 1.5 years to 2.5 years. We chose the latter because it is the date that obviously gets closer to the actual timing of the patented invention.

We said '*patents as point pattern*' as a result of a multivariate spatial point process which is a stochastic mechanism that generates different types of points in two-dimensional space. The qualitative feature that allow us to distinguish the different points is the industrial sector to which patents belong. Fig 1.1 gives the overall spatial distribution of the 12090 "*cases*". The map suggests a strong aggregation of cases around some *innovative* sub-regions: Milan's area and Triveneto regions are the more innovative ones. In general a large proportion of inventors live in the main cities' areas. Table 1.1 reports the distribution of patents among the six industrial sectors.

sector	N	patents' number	marginal proportion
ELECTRICITY ELECTRONICS	1	1358	0,11
INSTRUMENTS	2	1129	0,09
CHEMICALS PHARMACEUTICALS	3	893	0,07
PROCESS ENGINEERING	4	1915	0,16
MECHANICAL ENGINEERING MACHINERY	5	4373	0,36
CONSUMER GOODS CIVIL ENGINEERING	6	2422	0,20
TOTAL		12090	1

Table 1.1: Sectors distribution of patents in North-Italy

1.3 Point Processes

1.3.1 Point Process Theory

Stochastic point processes generate points in time, in space (2 dimensions) or in multidimensional space: we refer to these points as *events*. Cox & Lewis (1966,(11)) and Daley & Vere-Jones (1972,(13)) are the body of the theory of one-dimensional point

processes. The extension to two-dimensional space is due, among others, to Diggle (2003, (16)). In general we define a Stochastic Point Process, say X , in \mathbb{R}^d as:

- a random subset $X \subset \mathbb{R}^d$, locally finite, *i.e.* $X \cap A$ is finite \forall bounded region $A \subset \mathbb{R}^d$ (Borelian of \mathbb{R}^d)
- the measurability of X is equivalent to

$$N(A) \equiv \text{card}(X \cap A)$$

is a random variable \forall bounded region A , so we can refer to the process X as the count process $N(\cdot)$

- the distribution of X is characterized by the joint distribution of $N(A_1), \dots, N(A_n)$ $\forall A_1 \dots A_n$ disjoint regions and \forall integer $n \geq 1$

As from the classical theory of stochastic processes, a point process is called *stationary* if its distribution function in every region $A \subset \mathbb{R}^d$ is invariant to the translations of A ; it is called *isotropic* if the invariance holds for the rotations.

Every point process is characterized by first-order and second-order features. First-order features are described by the so-called *intensity function* (Diggle, 2003, (16)),

$$\lambda(a) = \lim_{|da| \rightarrow 0} \left\{ \frac{\mathbb{E}[N(da)]}{|da|} \right\}, \quad (a \in \mathbb{R}^d)$$

da is the infinitesimal region which contains the point a ; $|da|$ its area, or better the Lebesgue measure on this region. The *intensity function* provides informations on the mean number of events per unit area. Generally it varies over the space. In the particular case of a stationary process it assumes a constant value.

In the same way the second-order features are described by the *second-order intensity function* (Diggle, 2003, (16)):

$$\lambda_2(a, b) = \lim_{|da| \rightarrow 0, |db| \rightarrow 0} \left\{ \frac{\mathbb{E}[N(da)N(db)]}{|da||db|} \right\}, \quad (a \neq b; a, b \in \mathbb{R}^d).$$

As the previous, the *the second-order intensity function* varies over the space too, except for a stationary process where

$$\lambda_2(a,b) = \lambda_2(a - b);$$

and for an isotropic process, for which there aren't direction effects and, therefore, the dependence among points is due just to the distance:

$$\lambda_2(a,b) = \lambda_2(t)$$

where t is the distance between a and b (it often is the *Euclidean distance*). We can find information on the second-order features of a point process also in the *covariance density function* (Diggle, 2003, (16)), defined as

$$\gamma(a,b) = \lim_{|da|,|db| \rightarrow 0} \left\{ \frac{\mathbb{E}[\{N(da) - \lambda(a)\}\{N(db) - \lambda(b)\}]}{|da||db|} \right\}, \quad a \neq b, \quad a, b \in \mathbb{R}^d;$$

which is related with the *second-order intensity function*:

$$\gamma(a,b) = \lambda(a,b) - \lambda(a)\lambda(b).$$

We can extend this definition when $a = b$ assuming the process is *orderly*: i.e. at one point just one event can occur; this implies

$$Pr[N(da) > 1] = o(|da|)$$

and so

$$\mathbb{E}[\{N(da)\}^2] = \mathbb{E}[\{N(da)\}] = \lambda(a)|da|.$$

We can note that the *intensity function* λ describes the mean number of events occurring in a certain region. Whereas the *second-order intensity function* λ_2 characterizes

the mean number of additional events given a certain event.

If the *orderliness* holds, we can compute the *conditional intensity*. For a *orderly* point process

$$\mathbb{E}[N(dz)] - \text{Prob}\{N(dz) = 1\} = o(|dz|)$$

and

$$\mathbb{E}[N(dz)N(dz')] - \text{Prob}\{N(dz) = N(dz') = 1\} = o(\max(|dz|, |dz'|));$$

so we can compute the conditional density given a random variable:

$$\begin{aligned} \lambda(z|Y) &= \lim_{|dz| \rightarrow 0} \frac{\mathbb{E}(N(dz)|Y)}{|dz|} \\ &= \lim_{|dz| \rightarrow 0} \frac{\text{Prob}(N(dz) = 1|Y)}{|dz|}. \end{aligned}$$

Instead of the random variable, we compute the conditional density given that in z' an event occurs, we expect some changes:

$$\text{Prob}\{N(dz) = 1|N(dz') = 1\} = \frac{\text{Prob}\{N(dz) = N(dz') = 1\}}{\text{Prob}\{N(dz') = 1\}}.$$

So the conditional intensity:

$$\lambda(z|N(dz') = 1) = \lim_{|dz| \rightarrow 0} \left[\frac{\text{Prob}\{N(dz) = N(dz') = 1\}}{\text{Prob}\{N(dz') = 1\}} * \frac{1}{|dz|} \right];$$

and finally:

$$\begin{aligned} \lambda(z|\text{event in } z') &= \lim_{|dz|, |dz'| \rightarrow 0} \frac{\mathbb{E}[N(dz)N(dz')]}{|dz||dz'|} * \left[\frac{\mathbb{E}N(dz')}{|dz'|} \right]^{-1} \\ &= \lambda_2(z, z') \cdot \lambda(z)^{-1}. \end{aligned}$$

Just for point processes which are stationaries, isotropics and orderlies, Ripley (1977, (22)) gives a characterization of second-order features in terms of conditional intensity:

$$K(t) = \lambda^{-1} \mathbb{E}[\text{number of further events within a distance } t \text{ far from an arbitrary event}]$$

and known as the *K function*.

Starting from the previous hypothesis of *orderliness*, stationarity and isotropicity and using the previous results on the conditional intensity we have that:

$$\begin{aligned}\lambda K(t) &= \int_0^{2\pi} \int_0^t \{\lambda_2(a)/\lambda\} a \, da \, d\theta \\ &= 2\pi\lambda^{-1} \int_0^t \lambda_2(a)a \, da\end{aligned}$$

and the inverse relation:

$$\lambda_2(t) = \lambda^2(2\pi t)^{-1} K'(t)$$

1.3.2 Poisson Point Processes

The benchmark will be the process generating the CSR (*complete spatial randomness*): the Homogeneous Point Process. So, we could observe realizations, *pattern*, of such process, or also patterns which present more 'regularity' or more 'clustering' with respect of the CSR, these departures are captured by the previous tools.

Let be μ a local finite and diffuse measure defined on the regions in \mathbb{R}^d ; X is a *Poisson Point Process* with *intensity measure* μ if:

- (a) *independent scattering*: $N(A_1) \dots N(A_n)$ are independent for $A_1 \dots A_n$ disjoint regions and for integers $n \geq 2$
- (b) $N(A) \sim Po(\mu(A))$ for bounded regions A
- (a)-(b) are equivalent to (b)-(c)
- (c) $\forall A$ with $\mu(A) > 0$ and integer $n \geq 1$, the n points $x_1 \dots x_n$ in $X \cap A$, conditional to $N(A) = n$, are independent and each point has distribution in space

$$\frac{\mu(A \cap \cdot)}{\mu(A)}$$

Now we can return to the intensity function. If $\mu \ll l$, *i.e.* is absolutely continuous with respect to the *Lebesgue* measure, then it has density $\lambda : \mathbb{R}^d \rightarrow [0, \infty)$ so that $\mu(A) = \int_A \lambda(x)dx \forall A$. As consequence

$$\lambda(\xi) = d\mu(\xi)/d\xi$$

defined as *intensity function*.

The moments of $N(\cdot)$ for a Poisson process are easily computed

$$\mathbb{E}N(A) = \mu(A) \quad \text{and} \quad \text{Cov}(N(A), N(B)) = \mu(A \cap B).$$

Homogeneous Poisson Point Processes

As above a *stationary* and *isotropic* Poisson process is called *Homogeneous* and it will display a constant λ . Such process models the *CSR* (*Complete Spatial Randomness*).

Definition 1.3.1. Let be $\lambda \in \mathbb{R}_+$ fixed, the collection of events x_1, x_2, x_3, \dots in \mathbb{R}^d is a Homogeneous Poisson Process with intensity λ if:

1. For every bounded region $A \subset \mathbb{R}^d$ the number of events $N(A) \sim \text{Po}(\lambda\nu(A))$ where $\nu(\cdot)$ is the *Lebesgue measure*;
2. $\forall A \subset \mathbb{R}^d$ bounded region, conditional to the realization $N(A) = n$, the events x_1, x_2, x_3, \dots are an *i.i.d.* survey from a uniform distribution the region A .
(1) and (2) imply
3. $\forall A, B \subset \mathbb{R}^d$ disjoint, the R.V. $N(A)$ and $N(B)$ are independent.

Inhomogeneous Poisson Point Process

A general point process assigns its points in a given region with a bigger probability respect to other one due to the nature and to the features of the *intensity function*.

Definition 1.3.2. Let be $\lambda : \mathbb{R}^d \rightarrow \mathbb{R}_+$ any function, non-constant, on \mathbb{R}^d . The collection of events x_1, x_2, x_3, \dots in \mathbb{R}^d is a Inhomogeneous Poisson Process if:

1. For every bounded region $A \subset \mathbb{R}^d$ the number of events $N(A) \sim Po(\lambda(A))$, where $\lambda(A) = \int_A \lambda(x) dz$;
2. Conditional to the realization $N(A) = n$, the locations of events x_1, x_2, x_3, \dots are *i.i.d.* survey from the distribution on A with density proportional to $\lambda(x)$, $x \in A$.

The difference between Inhomogeneous and Homogeneous Poisson process is now evident: the mean number of events for unit of aerea varies because the intensity function varies in the first process and it is costant in the second one.

The Inhomogeneity could be due to the characteristics of the phenomenon under study and we can describe its nature modelling the intensity function using explanatory variables: a vector $Z(x)$ observed in the location x :

$$\lambda(x) = \lambda(Z(x)' \theta), \quad x \in \mathbb{R}^d, \quad \theta \in \mathbb{R}^{dim(\theta)} \text{ unknown parameters vector .}$$

So, variations of the *intensity function* lead to *clustering* or to *inhomogeneity*.

1.3.3 Cox Process

If the point process is not driven by a non-costant λ defined as above, but by a stochastic mechanism itself, the new point process is a *double stochastic* point process. The underlying stochastic mechanism captures esplicitly the spatial dependence among events.

Definition 1.3.3.

Let be $\{\Lambda(x) : x \in \mathbb{R}^d\}$

a stochastic process such that $\Lambda : \mathbb{R}^d \rightarrow \mathbb{R}$. The collection of events x_1, x_2, x_3, \dots in \mathbb{R}^d is a Cox Process driven by Λ if conditional to the realization of the underlying process, $\Lambda = \lambda$, the collection is a inhomogeneous Poisson Process with intensity λ .

1.3.4 Remarks: Inhomogeneous Poisson Process and Cox Process

A possible source of aggregation among events is environmental (or economics') heterogeneity. Specifically, an inhomogeneous Poisson process with intensity function $\lambda(x)$ will produce apparent clusters of events in regions of relatively high intensity. The source of such environmental heterogeneity might itself be stochastic in nature. This suggests investigation of a class of *doubly stochastic* processes formed as inhomogeneous Poisson processes with stochastic intensity functions. Such processes, as defined above, are called Cox processes, following their introduction in one temporal dimension by Cox (1995, (12)).

The Cox point process is stationary if and only if the intensity process $\Lambda(x)$ is stationary, and similarly for isotropy. A convenient and expressive terminology is to refer to the Cox process *driven by* $\Lambda(x)$.

First-order and second-order properties are obtained from those of the inhomogeneous Poisson process by taking expectations with respect to $\Lambda(x)$. Thus in the stationary case, the intensity is

$$\lambda = E[\Lambda(x)].$$

Also, the conditional intensity of a pair of events at x and y , given $\Lambda(x)$, is $\Lambda(x)\Lambda(y)$, so that

$$\lambda_2(x, y) = E[\Lambda(x)\Lambda(y)].$$

In the stationary, isotropic case this can be written as

$$\lambda_2(t) = \lambda^2 + \gamma(t),$$

where

$$\gamma(t) = Cov\Lambda(x), \Lambda(y)$$

and $t = \|x - y\|$. Note that, the covariance function $\gamma(t)$ of the intensity process is also the covariance density of the point process.

From a statistical viewpoint, the distinction between clustering and heterogeneity can only be sustained if additional information is available, for example in the form of covariates. Note that if we were able to model the intensity surface $\Lambda(x)$ through a regression equation in measured covariates, rather than as a realization of a stochastic process, the resulting point process model would become an inhomogeneous Poisson process.

Anyway in the absence of independent replication, a non-parametrically specified inhomogeneous Poisson process is indistinguishable from a Cox process. The equivalence of the two processes is established formally in Bartlett (1964, (4)).

1.4 Classical Spatial Analysis

Our first exploratory analysis' aims to identify the spatial structure of the patents' point pattern in order to give it an economical interpretation. To characterize a spatial point process we need to model his two features: namely the number of events in any planar region and where these points are located. For a general Poisson Process, let $\lambda(\mathbf{x})$ be a non-negative valued function over the plane, called the *intensity function*,

- for any planar region A , the number of events, $N(A) \sim Poisson(\mu(A))$ where
$$\mu(A) = \int_A \lambda(\mathbf{x}) dx$$

- given $N(A) = n$, the locations of the n events form an independent random sample from the distribution on A with pdf proportional to $\lambda(\mathbf{x})$.

More analytically, the intensity function is defined as the mean number of points for unit area:

$$\lambda(\mathbf{x}) = \lim_{|d\mathbf{x}| \rightarrow 0} \left\{ \frac{\mathbb{E}[N(d\mathbf{x})]}{|d\mathbf{x}|} \right\}$$

$d\mathbf{x}$ is the infinitesimal region containing point \mathbf{x} . $|d\mathbf{x}|$ is the area of that region.

The simplest spatial structure - the complete spatial randomness (CSR) - is modelled by the *homogenous poisson process* characterized by a constant, over space, intensity function $\lambda(\mathbf{x}) = \lambda \forall \mathbf{x}$. One departure from CSR is represented by the case in which points tend to form local concentrations, as it is displayed in the patents' case. This *aggregation* could be due to *clustering* in which points form functional groups strictly related each other; or due to *heterogeneity* described by a process in which the *intensity function* varies spatially, possibly due to environment's features. Empirically we can't distinguish these two types of aggregation (Bartlett, 1964 (4), see also section 1.3.4 for further details). Both types of aggregation could be modelled by the so called *inhomogenous poisson process* which presents an intensity function $\lambda(x)$ that varies spatially giving to regions different probabilities to receive an event - a patent in our case. It seems natural to assume for the patents distribution an inhomogeneous Poisson point process, which can model the aggregation's feature of patents data.

Starting from the data plotted in Fig 1.1, we are able to make inference on the intensity of the spatial point process of patents in North-Italy, said region A , which characterizes the whole spatial distribution. We used a 2-dimensional non-parametric kernel estimator. Dealing with Kernel estimation in two dimensions we can imagine a window, a moving three-dimensional function called kernel, which weights events within its sphere of influence according to their distance from the point where the intensity is estimated.

The method is commonly used in a more general statistical context to obtain smooth estimates of univariate (or multivariate) probability densities from an observed sample of observations (Silverman 1986, (23); Wand and Jones 1995, (26)). Estimating the intensity of a spatial point pattern is similar to estimating a bivariate probability density, as follows

$$\hat{\lambda}_h(x) = \frac{1}{p_h(x)} \sum_{i=1}^n \frac{1}{h^2} \mathbf{K} \left(\frac{x_i - x}{h} \right)$$

where $x_i, i = 1, 2, \dots, n$ are observed data locations and $p_h(x) = \int_A \frac{1}{h^2} \mathbf{K} \left(\frac{x_i - x}{h} \right)$ is the edge correction factor for location x (Diggle 1985, (15))

Here, $K()$ represents the kernel weighting function which, for convenience, is expressed in standardized form and 'stretched' according to the parameter h which is referred to as the *bandwidth*; the value of h is chosen so as to provide the required degree of smoothing in the estimate. As to the exact functional form of the kernel, $K()$, we require a decreasing radially symmetric bivariate function providing a total weight of unity over the region of influence. Different choices from amongst the range of 'reasonable' candidates have relatively little effect on the resulting intensity estimate. Typical choices might be the Gaussian Kernel and the Quartic Kernel.

The kernel estimate is sensitive to the choice of bandwidth, h . As this increases, there is more smoothing of the spatial variation in intensity; as it is reduced we obtain an increasingly 'spiky' estimate. What value, then should we choose? In practice, the value of kernel estimation is that one that has the flexibility to experiment with different values of h , exploring the surface $\hat{\lambda}_h(x)$ using different degrees of smoothing in order to look at the variation in $\lambda(x)$ at different scales. There are also methods which attempt automatically to choose a value of h which optimally balances the reliability of the estimate against the degree of spatial detail that is retained, given the observed pattern of event locations (Diggle, 1985); some of them will be used in the following

sections.

Considering just the North of Italy (see Table 1.1), the most innovative and the most interesting region to study, we can estimate, as described above, and plot the intensity for the total point pattern of patents and for each component pattern (one for each of the six economic sectors). The choice of the optimal bandwidth is reached minimizing the MSE as in Berman & Diggle (1989, (5)). Seeing at results in Fig. 1.1 and in Fig. 1.2 and in Fig. 1.3 (blue points represent some main cities), the clusters detected in the total process are observed also in each component process indicating non real sector-district. Aggregation around Milan and Turin and in south-Triveneto that we observe in any sector are due just to the presence of a very intensive innovation production and not to a particular sector activity.

1.4.1 Some notes on how I chose the optimal bandwidth

Recall that one definition of the K-function of a stationary process is that

$$K(t) = 2\pi\lambda^{-2} \int_0^t \lambda_2(s)s \, ds \quad (1.4.1)$$

where λ and $\lambda_2(s)$ are the intensity and the second-order intensity, respectively. It turns out that several non-parametric inference problems for point process data can be solved by estimating a weighted integral,

$$K_\phi(t) = 2\pi\lambda^{-2} \int_0^t \phi(s)\lambda_2(s)s \, ds, \quad (1.4.2)$$

for suitably defined, problem specific functions $\phi(s)$. Before considering the specific application, we give the general results from Berman & Diggle (1989, (5)). It is convenient to re-express (1.4.2) as

$$K_\phi(t) = 2\pi\lambda^{-2} J(t), \quad (1.4.3)$$

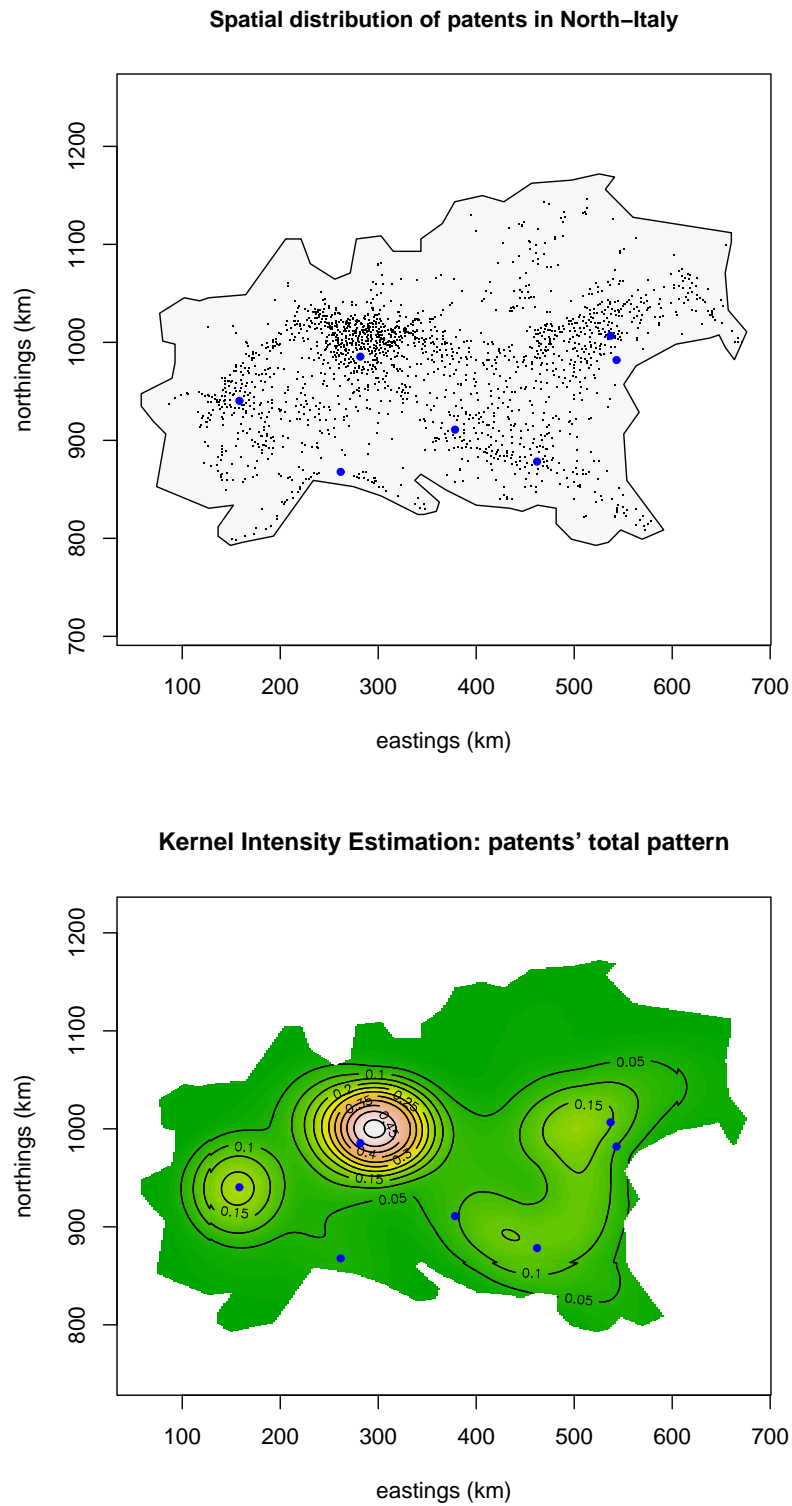


Figure 1.1: a) Spatial Distribution of all patents in ninety b) Kernel Intensity Estimation: All Patents Pattern

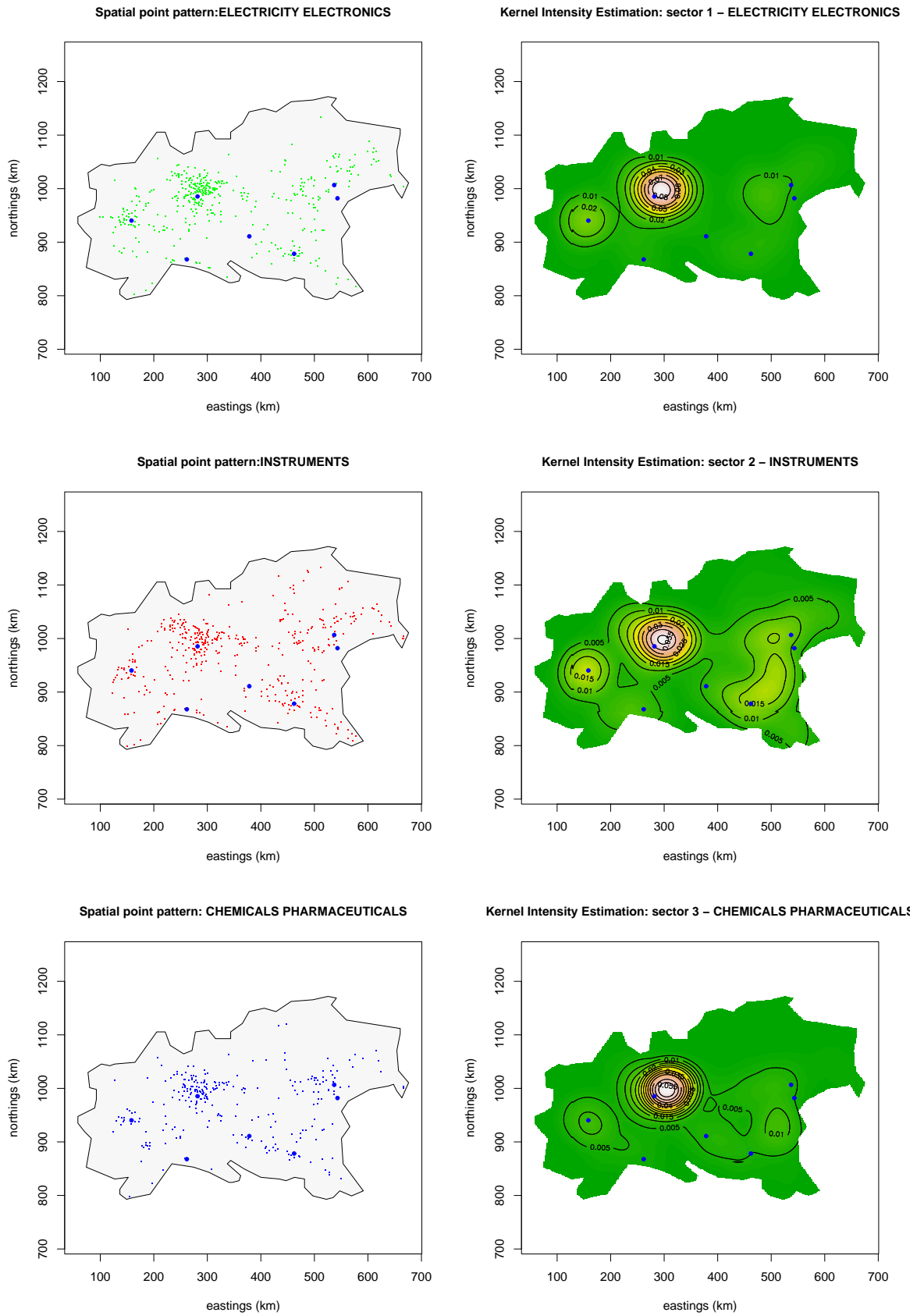


Figure 1.2: Kernel Intensity Estimation: Single Sectors

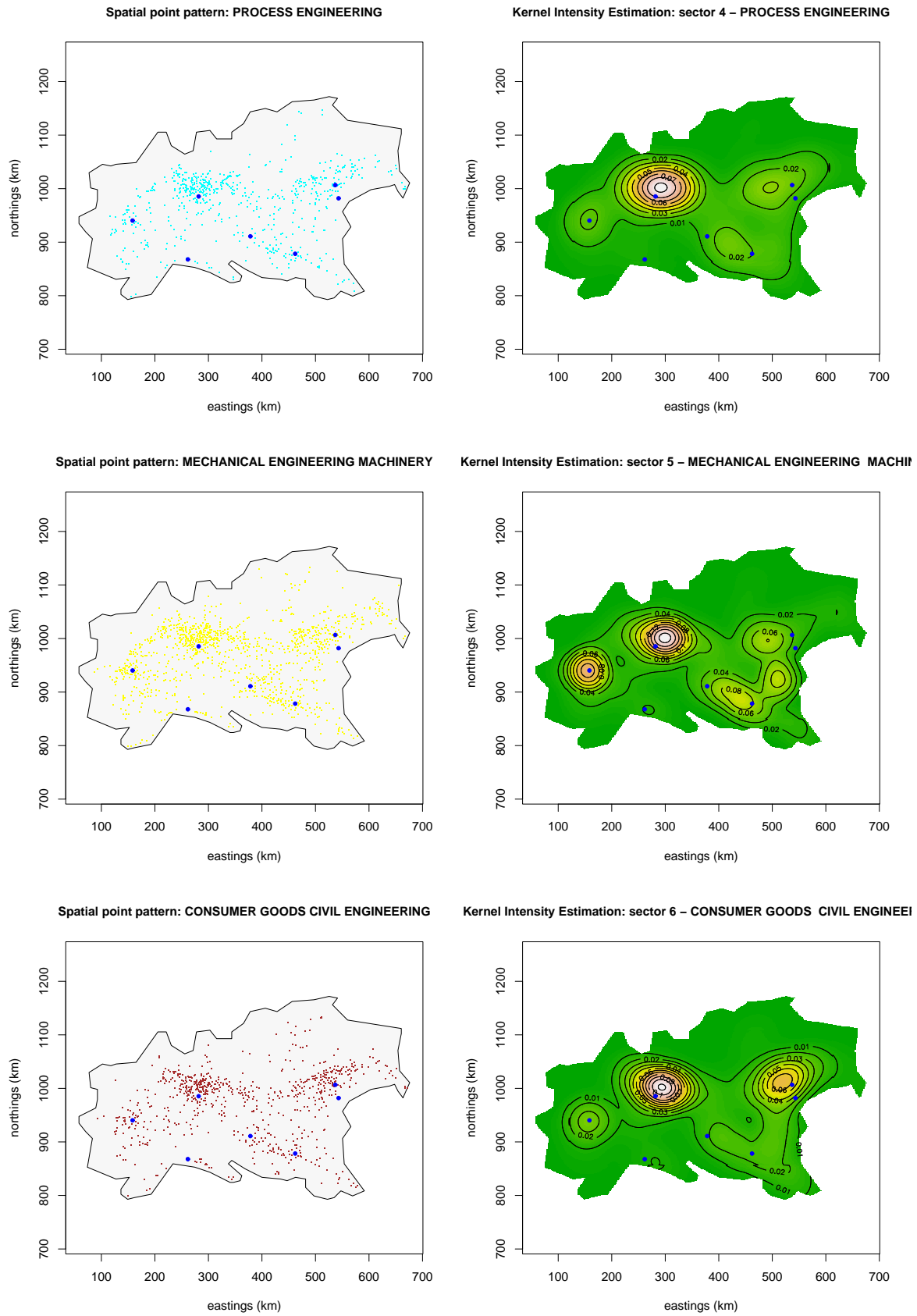


Figure 1.3: Kernel Intensity Estimation: Single Sectors

where

$$J(t) = \int_o^t \phi(s)\lambda_2(s)s ds. \quad (1.4.4)$$

Using integration by parts to evaluate (1.4.3), and substituting from (1.4.1), we obtain

$$J(t) = \lambda^2(2\pi)^{-1} \left(K(t)\phi(t) - \int_o^t K(s)\phi'(s)s ds \right). \quad (1.4.5)$$

Estimation of $J(t)$ is now straightforward, because $\phi(\cdot)$ is a known function and we can substitute existing estimators for λ and for $K(t)$ into (1.4.5). In practice, the integration on the right-hand side of (1.4.5) must be carried out numerically, but this is usually straightforward and numerically stable.

Diggle used this result to derive and estimate the mean square error on the assumption that the underlying process is a stationary, isotropic Cox process (remember that, under some assumptions the inhomogeneous Poisson process and the Cox process are indistinguishable). If the driving intensity of a Cox process has expectation λ and covariance function $\gamma(u)$, then the Cox process itself has intensity λ and second-order intensity $\lambda_2 = \gamma(u) + \lambda^2$. Let $N(x, h)$ denote the number of events of the Cox process within distance h of the point x . Then, temporally ignoring edge effects, the non-parametric estimator of the realized value of $\Lambda(x)$ described above can be written as

$$\tilde{\lambda}(x) = N(x, h)/(\pi h^2). \quad (1.4.6)$$

We now consider the mean square error of $\tilde{\lambda}(x)$,

$$MSE(h) = E[\tilde{\lambda}(x) - \Lambda(x)]^2. \quad (1.4.7)$$

where the expectation is with respect to the distribution of the Cox process, i.e. with respect both to $\Lambda(\cdot)$ and to the points of the process conditional to $\Lambda(\cdot)$. After some considerations and manipulations, Diggle re-writes the (1.4.7) as follows:

$$MSE(h) = \lambda_2(0) + \lambda^2 - 2\lambda K(h)/(\pi h^2) + (\pi h^2)^{-2} \int \int \lambda_2(\|x - y\|) dy dx. \quad (1.4.8)$$

The first term on the right-hand side of (1.4.8) does not depend on h , and it follows that the value of h which minimizes $MSE(h)$ also minimizes

$$M(h) = \lambda - 2\lambda K(h)/(\pi h^2) + (\pi h^2)^{-2} \int \int \lambda_2(\|x - y\|) dy dx. \quad (1.4.9)$$

Now, of the two terms on the right-hand side of (1.4.9), one is an explicit function of $K(h)$, which can be estimated by substituting the standard estimator \widehat{K} , whilst the double integral can be converted to a single integral of the form (1.4.2) using polar coordinates, and can therefore be estimated as described above.

To introduce the edge effects we just to use as the estimator for $\widehat{\lambda}$ describe in the previous section (Diggle, 1985, (15)).

In order to decide which value of h has to be used in the non-parametric kernel estimator described above, we must just minimize the function M (1.4.9). This framework has been implemented by Diggle and Rowlingson in the *SPLANCS* package available for the *R* software.

1.5 Spatial Segregation

1.5.1 Spatial Segregation Theory

We deal with multivariate data and our interest falls in detecting the differences among behaviours of the spatial patterns of different industrial sectors. More precisely, we want to estimate the spatial segregation in the spatial distribution of patents' data: if some particular type of points, typified by different industrial sectors, predominant in one particular subregions more than in others, we say that segregations occurs. More precisely a multivariate pattern exhibits spatial segregation if the conditional intensity of type j points at x given a point of type i at x is less than the marginal intensity of type j point at x (Diggle et al., 2005 (17)).

The multivariate inhomogeneous Poisson point process hypothesis made above assumes

for the component processes, one for each industrial sector, independent Poisson process distribution with intensity functions respectively $\lambda_k(x) : k = 1, \dots, m$ where k denotes the industrial sector.

As in Diggle et al. (2005, (17)), $p_k(x)$, the *type-specific probabilities*, denotes the conditional probability that a case known to occur at location x is of type k :

$$p_k(x) = \frac{\lambda_k(x)}{\sum_{j=1}^m \lambda_j(x)}.$$

This tool allows us to study spatial segregation. A constant type-specific probability, $p_k(x) = p_k$, means a completely unsegregated underlying Poisson process. It also means that different typologies of points show no propensity to occur in a particular subregion. Instead, in the opposite case of complete spatial segregation, for each location x , $p_k(x) = 1$ for some type and $p_k(x) = 0$ for the others, i.e. at every location, only one type of point occurs. Between these two extreme situations, there are various forms of spatial segregation which can be expressed by the spatial distribution of the estimated functions $\hat{p}_k(x) : k = 1, \dots, m$. A statistical framework to estimate the type-specific probabilities was proposed by Diggle et al. (2005, (17)), in which a kernel regression estimator is used for probabilities surfaces $p_k(x)$ as follows: the data are represented as multinomial outcomes $Y_i, i = 1, \dots, n$, where $Y_i = k$ denotes a patent belonging to k industrial sectors, so:

$$\hat{p}_k(x) = \sum_{i=1}^n w_{ik}(x) I(Y_i = k),$$

where, for each industrial sector k ,

$$w_{ik}(x) = w_k(x - x_i) / \sum_{j=1}^n w_k(x - x_j),$$

with $w_k(\cdot)$ a kernel function with band-width $h_k > 0$, thus $w_k(x) = w_0(x/h_k)/h_k^2$, where $w_0(\cdot)$ is the standardised form of the Gaussian kernel function: $w_0(x) = \exp(-\|x\|^2/2)$.

The log-likelihood function is

$$L(p_1, \dots, p_m) = \sum_{i=1}^n \sum_{k=1}^m I(Y_i = k) \log p_k(x_i), \quad (1.5.1)$$

where $I(\cdot)$ is the indicator function. In a parametric model for p_k , a widely accepted method of parameter estimation is to choose parameter values to maximise the right-hand side of the previous equation 1.5.1. In the kernel setting, doing so would lead to the unhelpful band-width choices $h_k = 0$, giving $\hat{p}_k(x_i) = 1$ or 0 according to whether the corresponding Y_i is equal to k or not. To circumvent this problem, we use a cross-validated log-likelihood function. We select a common band-width for all sector and it is chosen maximising the cross-validated log-likelihood:

$$L_c(h) = \sum_{i=1}^n \sum_{k=1}^m I(Y_i = k) \log \hat{p}_k^{(i)}(x_i),$$

where $\log \hat{p}_k^{(i)}(x_i)$ denotes the proposed kernel estimator based on all data except (x_i, Y_i) .

1.5.2 Application and Tests proposed

The first impression we have, watching the spatial point pattern of inventors locations in Fig. 1.1 and in Fig. 1.2 and in Fig. 1.3 for each of the six industrial sectors, is that no strong segregation occurs. They seem to have a quite common distribution, but we will discover that this impression is wrong.

The statistical tools described above allow to check our guess. Firstly, we must choose the optimal band-width for the proposed kernel estimator. Figure 1.4 shows the cross-validated log-likelihood and suggests to set an optimal value $h_{opt} = 9Km$. In such a way we can read the obtained \hat{p}_k -surfaces with their peaks and troughs, and

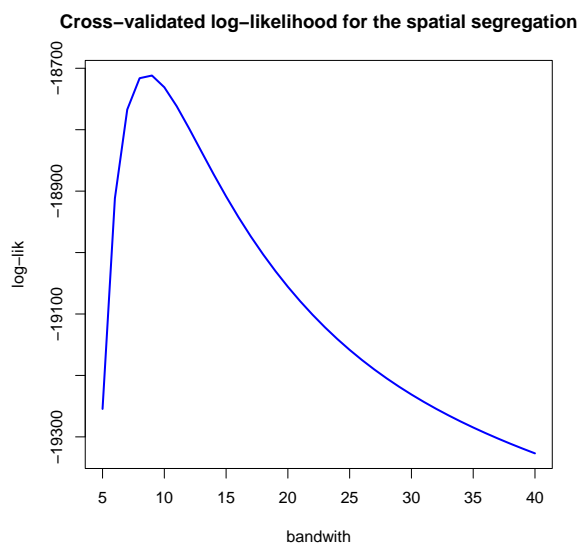


Figure 1.4: Cross-validated log-likelihood for the six industrial sectors

assess if there is spatial segregation. However we also want to test how many of them are statistically significant: some of them could be quite surprising. For this reasons we implement a Monte Carlo *local spatial test*: for each point of a given grid covering the studied region we estimate the type-specific probabilities for each sector. Then, we perform 99 simulations, re-labelling the data at random whilst preserving the observed number of cases of each type; for each simulated dataset we thus compute the estimated type-specific probabilities over the given grid. In this way, we obtain the p-value of each estimate at each point of the grid, enabling us to plot only the statistically significant estimated probabilities in order to detect if there were some *strange results* to be reworded. The new maps are plotted in Fig.1.5 and Fig.1.6 and Fig.1.7. Estimation has been carried out using R software (www.R-project.org) with the special package *spatialkernel* written by Pingping Zheng. White zones represent omitted unreliable estimates. We have to admit that most estimates are statistically significant. Now a strong segregation is evident in some subregions for some sectors.

In order to detect possibly presence of segregation, we compare, for each sector, the estimated type-specific probability maps to the corresponding constant theoretical probability of no segregation (*the marginal proportions*), which are reported in Table 1.1. If we don't expect segregation in a subregion for a given sector, then the type-specific probability must be equal to the marginal proportion of that sector. So, for instance, in the sector 1 map in Fig. 1.5 where the estimate exceeds the marginal proportion of 11,23% (which represents the probability of no segregation), there we can detect a real district of the sector 1 *electricity electronics*. Instead, where it is lower than the benchmark, we have to consider a weak innovation's power for the considered sector. In such a way, interesting considerations could be made. *Electricity and Electronics* sector shows innovative areas inside the quadrilateral Milan, Turin, Genoa, Parma; furthermore Genoa's areas hosts a district of *Instruments* as the north part of Triveneto. It's interesting to underline a tongue of innovation for the *Chemicals&Pharmaceuticals* sector starting in the south of Turin, moving between Milan and Genoa and ends between Milan and Parma. A large district for the Process Engineering sector is detected in the middle of north Italy around Milan covering almost the whole area of Lombardia. For this sector, small but intensive districts are detected near Genoa and around Bologna. The *Mechanical Engineering Machinery* sector shows a very strong segregation and a very powerful district in Emilia along the line Bologna - Parma. It is a well-known motor district. Piemonte too seems to show a diffuse concentration of innovation activity in this sector. Consumer goods and Civil Engineering sectors aggregate in Veneto and in a very small district inside the triangle Turin, Genoa and Milan.

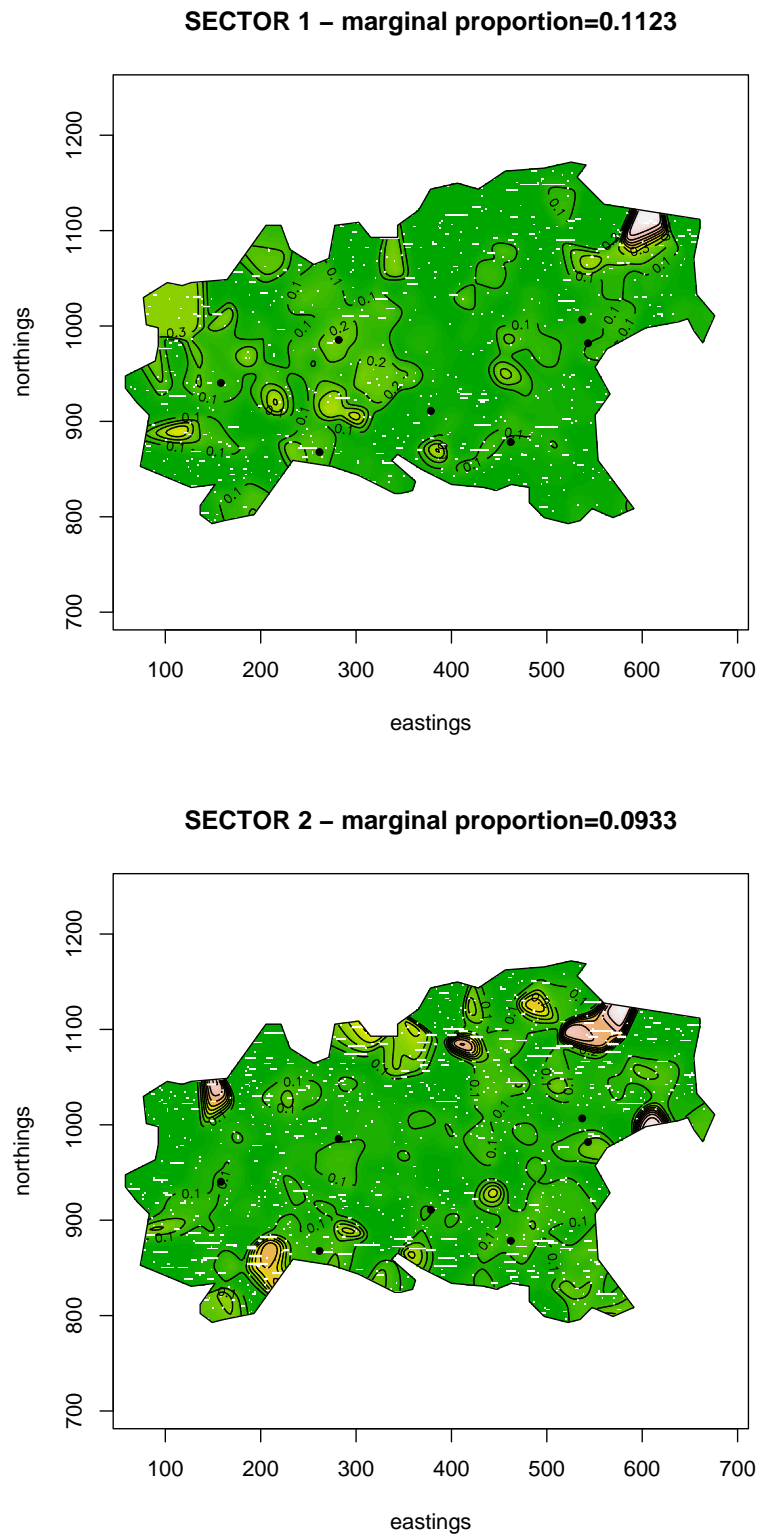


Figure 1.5: Statistically Significant Estimated type-specific probabilities

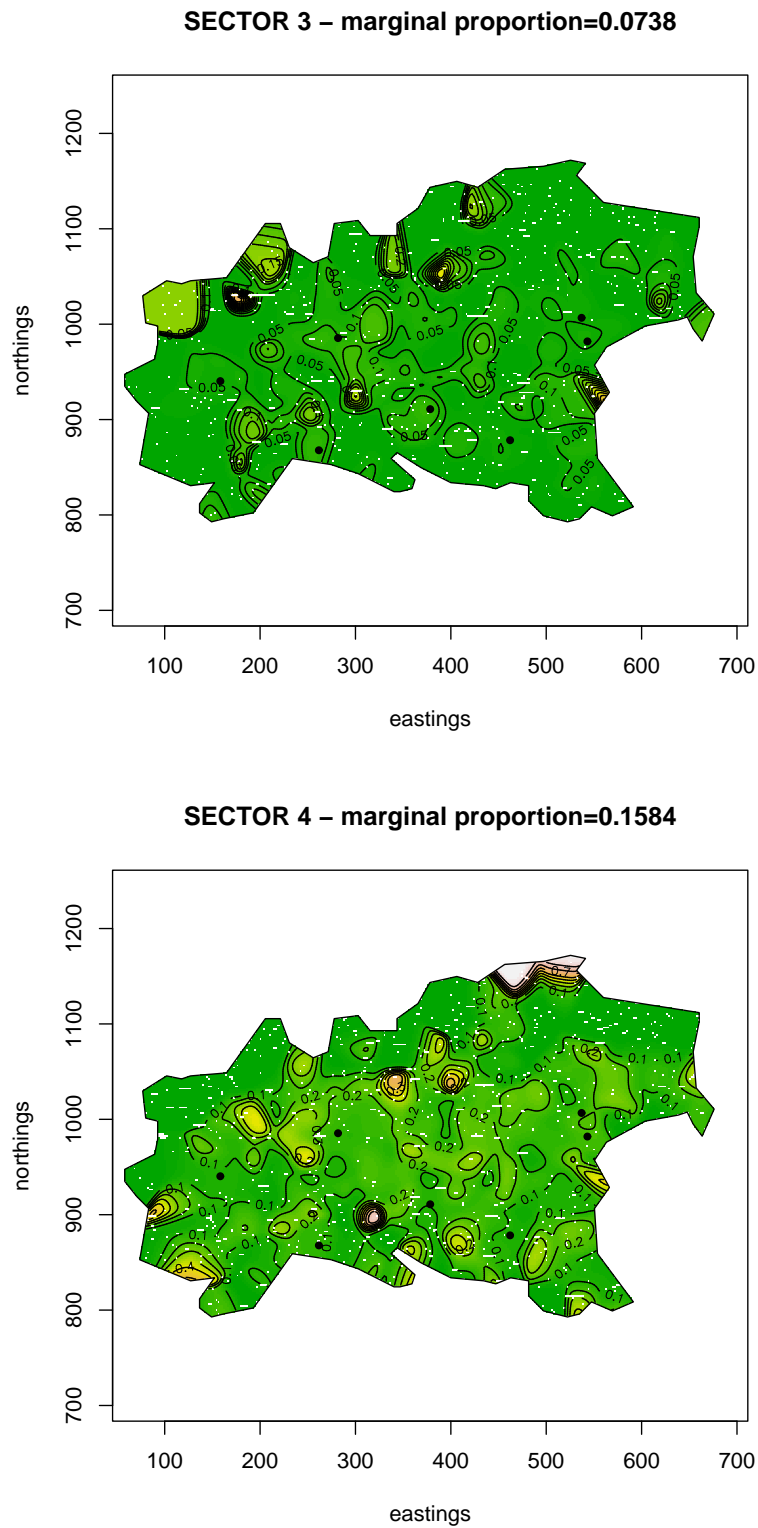


Figure 1.6: Statistically Significant Estimated type-specific probabilities

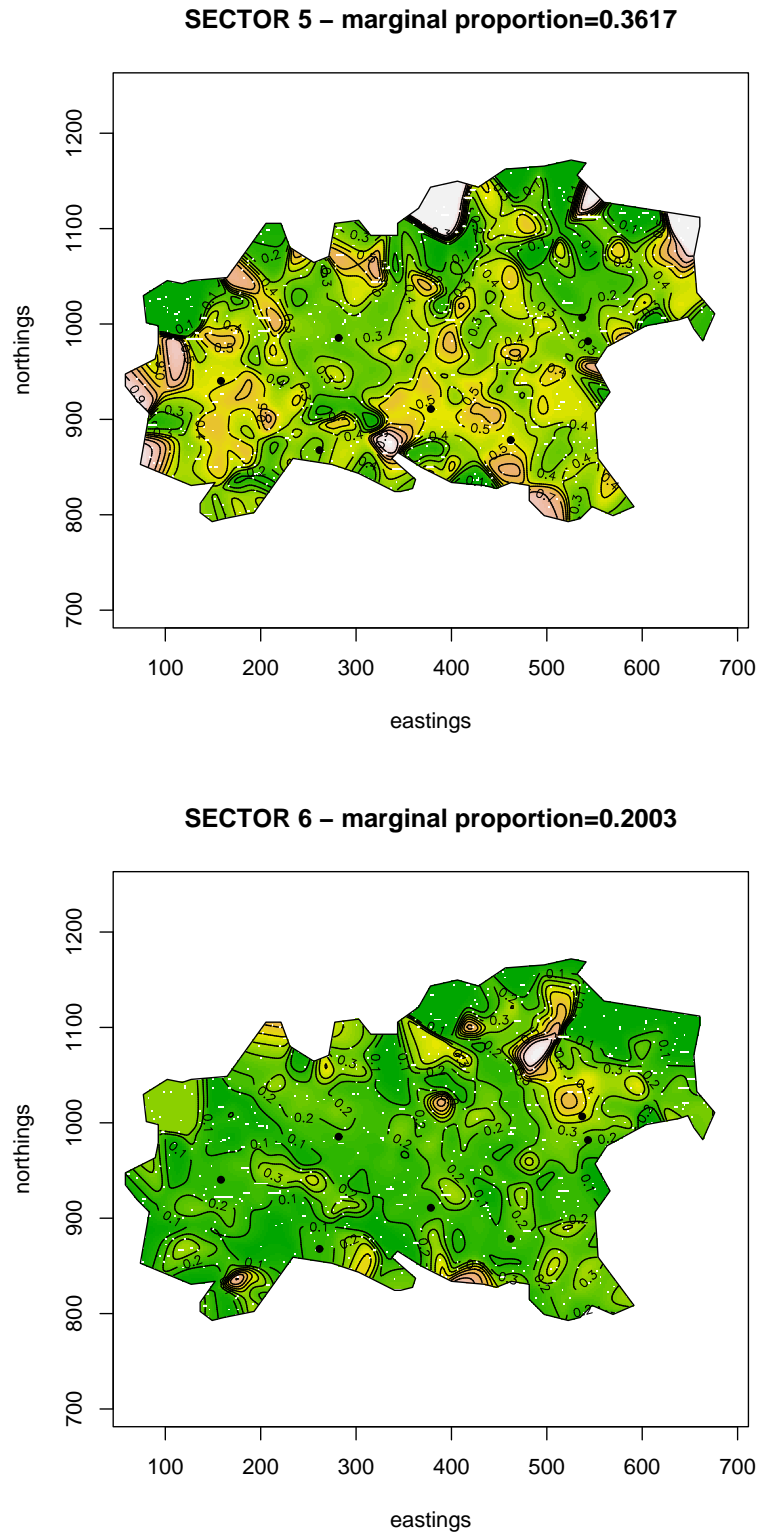


Figure 1.7: Statistically Significant Estimated type-specific probabilities

1.5.3 Monte Carlo Global Test

Diggle et al. (2005, (17)) proposed *the Global spatial test*, a Monte Carlo method to test directly the null hypothesis of no-segregation: $H_0 : p_k = \alpha_k$ for all x_i , i.e. constant type-probabilities. The estimator $\hat{\alpha}_k = n_k/n$ is proposed, where n_k is the number of cases of type k and n is the total number of cases. So the test statistic proposed is

$$T = \sum_{i=1}^n \sum_{k=1}^m \{\hat{p}_k(x_i) - \hat{\alpha}_k\}^2.$$

The value of T for the data t_1 is compared to the value of T, t_2, \dots, t_s , for new datasets obtained via Monte-Carlo simulation under H_0 , re-labelling the data at random whilst preserving the observed number of cases of each type. The p -value for this test is $p = (q + 1)/s$, where q is the number of $t_j > t_1$.

The Global Monte Carlo test described above, rejects the null hypothesis of no segregation with a p -value equal to 0.01: we already supposed this result seeing at maps and now it is confirmed.

1.5.4 Conclusions

Intensities estimation of each component point process starting from a multivariate point pattern could lead to detect non interesting clusters just because they should reflecting the clusters present in the total multivariate point pattern. We solved this problem analysing the spatial segregation of the patents distribution in north-Italy. So a kernel regression estimator was applied to estimate the type-specific probabilities, we said the conditional probability that a patent known to occur at location x is of the sector k . In such a way, many sector districts, which were previously hidden, have been detected. A Monte Carlo statistical test against the null hypothesis of no segregation was also provided.

Obtained empirical results have been very interesting, since they have diminished the

importance of the Milan's area, which seemed, from the classical spatial analysis, to be the only source of innovation in Italy. And, more important, these results have discovered explicitly some industrial districts very known among experts, as the motor-district between Bologna and Modena (firms as Ferrari, Maserati, Lamborghini, Ducati, ect.), as the Instruments district around Genoa, as the concentration of Process Engineering innovation activity in the whole Lombardia.

1.6 Temporal Analysis

1.6.1 Monte Carlo Test: same distribution among sectors over time?

We want to investigate the temporal behaviour of the patents' production, with a particular attention to the differences among different sectors. What we can see is a quasi-common temporal pattern: all sectors show an increase of production's rate from mid nineties to the end of nineties, except the *chemicals&pharmaceuticals* sector that displays the fall of its production's rate in the last two years after a very big expansion in mid-nineties middle. Empirical cumulative distribution function for each sector is plotted against the uniform distribution (Fig:1.8): the slope of each curve suggests us the production rate of that period. It is crucial to test the behaviour of the temporal pattern with respect the null hypothesis of same distribution over time, which could exclude any economical interpretation.

We construct a Monte Carlo test to check the null hypothesis of uniform distribution for all sectors; due the presence of ties we can't use the Kolmogorov-smirnov test. We compute the empirical cumulative distribution function - eased by the big size of dataset - for each sector $\hat{\mathbb{F}}_k(t)$, $k = 1, \dots, 6$ and compare it to the uniform's one $\mathbb{F}_U(t)$

computing the test statistic

$$T = \sum_{k=1}^6 \sum_{t=1}^n (\hat{\mathbb{F}}_k(t) - \mathbb{F}_U(t))^2 w_k$$

where t is the time (days) and the weights $w_k = \frac{\text{number of patents in sector } k}{\text{number of all patents}}$. We don't know the distribution of this statistics, so we implement a Monte Carlo simulation with a random re-labelling of the sector-index k whilst preserving the observed patents' number of each sector. We call t_1 the value of the statistic of the original data and t_2, t_3, \dots, t_m the values of the statistic after each random re-labelling. In this way, the p-value of significance for our Monte Carlo test is $p = (q + 1)/m$ where q is the number of $t_j > t_1$.

In this study, we ran the Monte Carlo test above with m , number of simulation equal to 999. The result is a p-value equal to 0.003 that allows us to reject the null hypothesis of same distribution over time of the patents, contradicting our first impression. We could ask ourself if this result is much influenced by the behavior of sector 3 - Chemicals&Pharmaceuticals. So we perform the same Monte-Carlo test for all sectors except sector 3. This produces as result a p-value equal to 0.018. Also in this case we can't say that sectors belongs to same temporal distribution, although the effect of removing *Chemicals&Pharmaceuticals* sector is sensible.

1.6.2 Temporal Segregation

Although we can't say that the different temporal patterns belong to same distribution, it could be much more interesting to extend the segregation analysis to the temporal dimension: dealing now with a pure temporal multivariate pattern, *segregation* occurs when one or more types of points predominate in particular time-periods more than

in others. We make the same main hypothesis for the model: a multivariate inhomogeneous Poisson point process, now in its temporal version in which each component process has a temporal intensity function which varies over time,

$$\lambda_k(t) : k = 1, \dots, m$$

where k denotes the industrial sector and $t \in (0, \infty)$, is a time-point.

The objects of our analysis: the **type-specific probabilities** become the conditional probabilities that a case, known to occur at time t , is of type k :

$$p_k(t) = \frac{\lambda_k(t)}{\sum_{j=1}^m \lambda_j(t)}.$$

A constant type-specific probability, $p_k(t) = p_k$, means temporal unsegregation for the underlying process. It is natural, in a temporal setting too, to test the null hypothesis of no-segregation via a time-adapted version of Monte-Carlo test built for the spatial model. The test statistic

$$S = \sum_{i=1}^n \sum_{k=1}^m \{\hat{p}_k(t_i) - \hat{\alpha}_k\}^2$$

is computed for the original dataset and for 99 simulated datasets obtained just relabelling the sector-index. The rank of S computed at the original data relatively to that computed at simulated datasets gives us the p-value of test.

Estimating methodology for temporal type-specific probabilities is the same used for spatial analogues: the kernel regression estimator,

$$\hat{p}_k(t) = \sum_{i=1}^n w_{ik}(t) I(Y_i = k),$$

where for each industrial sector k ,

$$w_{ik}(t) = w_k(t - t_i) / \sum_{j=1}^n w_k(t - t_j),$$

and w_k is a Gaussian kernel function whose optimal band-width, common for all sectors, h is chosen maximizing the cross validated log-likelihood as in spatial setting.

The optimal band-width for our dataset is 135 days (see fig. 1.9).

The *global temporal Monte Carlo test* gives a p-value equal to 0.01, which allow us to reject strongly the null hypothesis of temporal unsegregation. So we want to investigate and estimate the *type-specific temporal probabilities* in order to understand in a better way the differences among sectors. For each sector we plot (Fig. 1.10) the obtained estimates and their average, both in a gray color and with different colors only the estimates that are statistical significant ($p - value > 0.975$ or $p - value < 0.025$).

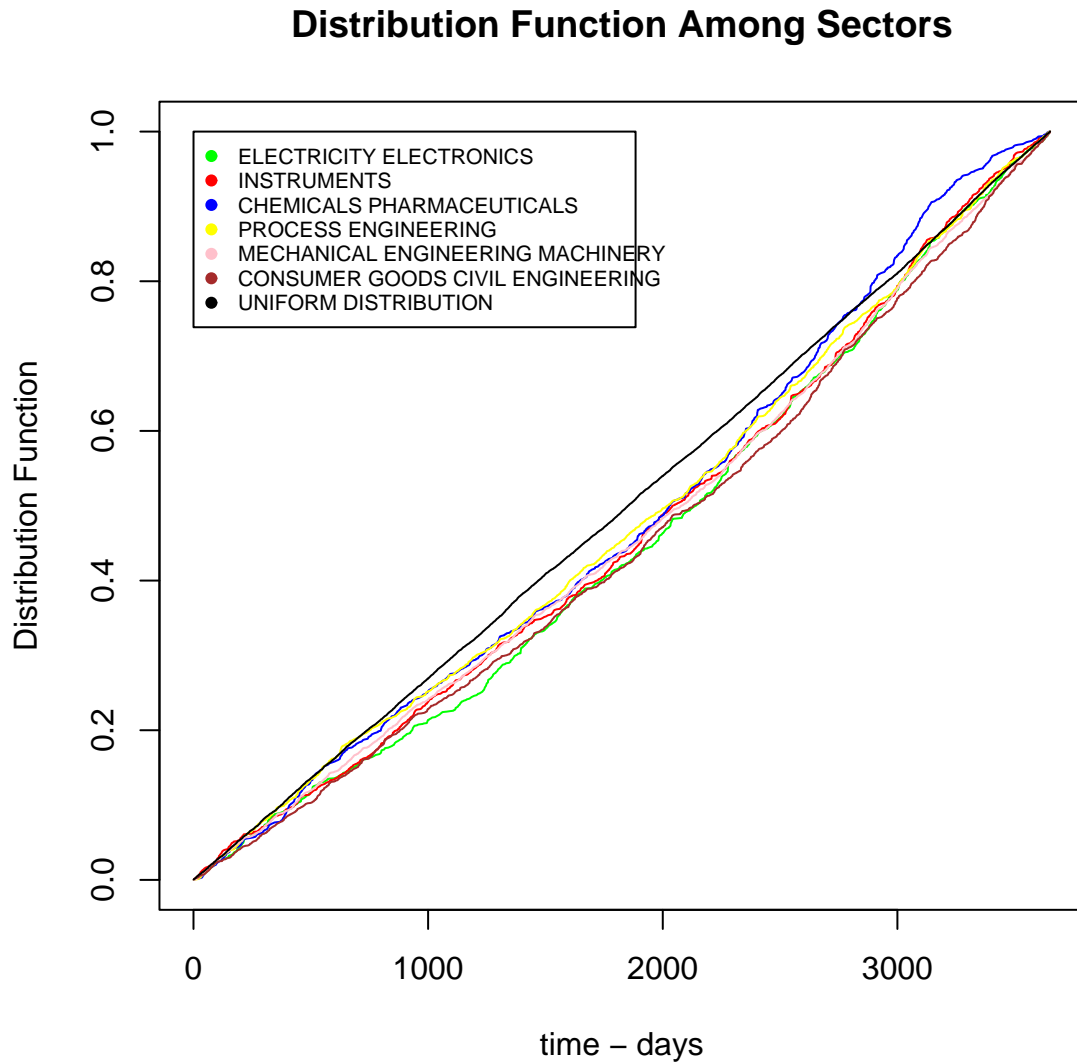


Figure 1.8: Empirical Distribution Function among Sectors compared to the Uniform Distribution

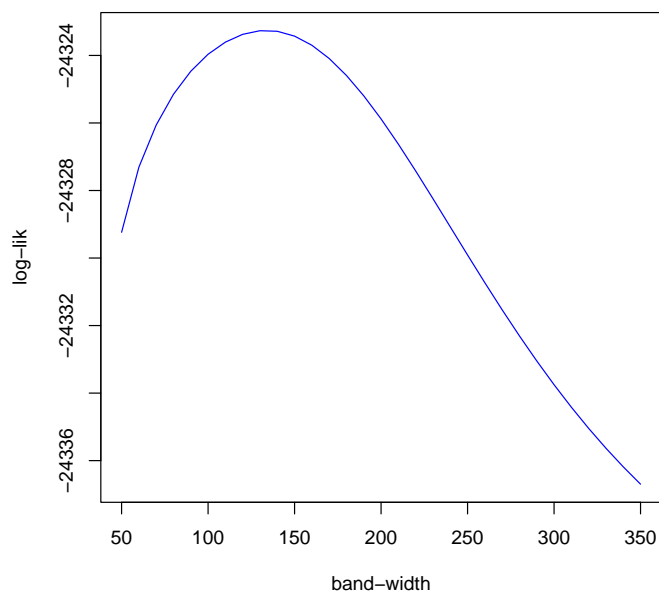


Figure 1.9: Temporal Cross-Validated Log-Likelihood

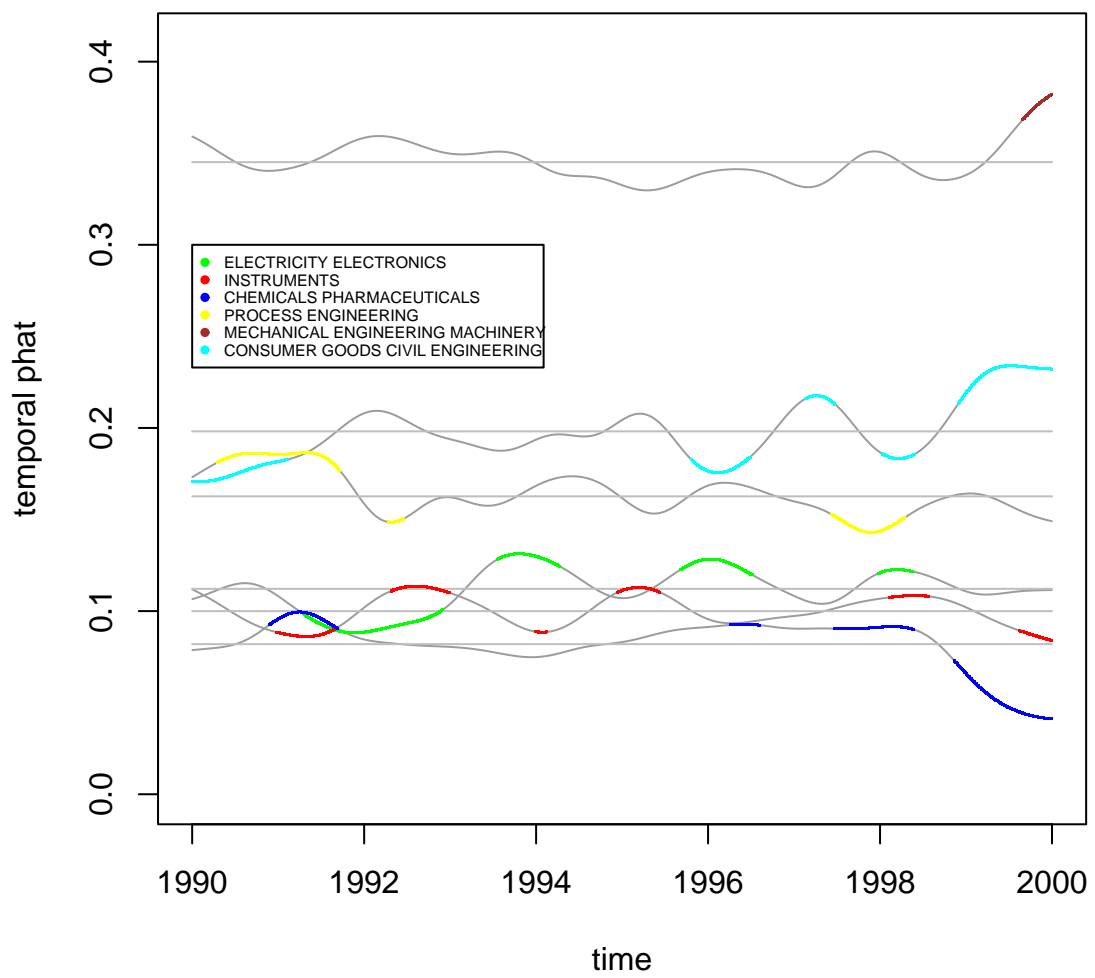


Figure 1.10: Temporal Type-Specific Probabilities

1.6.3 Intensity Kernel Estimator

We supplied a tool capturing the temporal features of each sector pattern relative to the others. We need now a tool that characterizes the absolute temporal pattern for each single sector.

We made the temporal Poisson processes hypothesis, so we have to estimate their intensity functions, i.e the mean number of events in a temporal unit. The statistical framework is the same used in Diggle (1985, (15)) that provides the kernel estimator:

$$\hat{\lambda}_k(t) = \frac{\sum_{i=1}^n \frac{1}{h} \omega\left(\frac{t-t_i}{h}\right)}{\int_0^T \frac{1}{h} \omega\left(\frac{t-u}{h}\right) du}, \quad k = 1 \dots 6 (\text{sectors})$$

where ω is the kernel function with bandwidth h which determines the amount of smoothing, $t_1 \dots t_n$ are the time-events. Note that the numerator is a standard intensity kernel estimator, instead the denominator is its edge-correction. A Gaussian kernel function is used, so we have:

$$\hat{\lambda}_k(t) = \frac{\sum_{i=1}^n \frac{1}{h\sqrt{2\pi}} \exp\left\{-\frac{(t-t_i)^2}{2h^2}\right\}}{\int_0^T \frac{1}{h\sqrt{2\pi}} \exp\left\{-\frac{(t-t_i)^2}{2h^2}\right\} dt}$$

It is well recongnized (Bowman&Azzalini 1997 (6), Diggle 1985 (15), Wand&Jones 1995 (26)) that the choice of the bandwidth h is far more important than the choise of the kernel's form. A variety of methods exist with this aim. We used a Cross-Validation criterium to choose the optimal smoothing parameter which is the minimizer of the following:

$$CV(h) = \int_0^T \hat{\lambda}_h^2(t) dt - 2 \sum_{i=1}^n \hat{\lambda}_h^{-i}(t_i)$$

where

$$\hat{\lambda}_h^{-i}(t_i) = \sum_{j=1; j \neq i}^n w_h(t_i - t_j)$$

is the so called *leave-one-out* intensity estimator. Mathematically the results of this procedure are accurate, but in our case they give a very small parameter. So we prefer

to make an heuristic decision: due to economic considerations, we choose a bandwidth equal to four months: this window's width looks to be natural to capture the temporal pattern's features which we are interested in. It's not reasonable using a small window: it's not so important understand the variations on a very small temporal scale. The results for each sectors are reported in Fig. 1.11, in y-axis the unit is the number of patents for each year.

1.6.4 Conclutions

To better understand the temporal behaviour and to make a comparison among sectors, we plotted in Fig.1.12, Fig.1.13, Fig.1.14, Fig.1.15, Fig.1.16 and Fig.1.17, for each industrial sector, the estimated temporal type-specific probabilities (*the relative pattern*) and the estimated intensity function of the underlying temporal Poisson process (*the absolut pattern*).

Peaks in each sector's patents' production in absolute value, captured by the kernel intensity estimation (on the rigth of Fig. from 1.12 to 1.17) sometimes became less important if we consider the temporal segregation captured by the *temporal type-specific probabilities* (on the left of Fig. from 1.12 to 1.17) maybe due to the fact that in the period considered all the sectors have seen an increase of their innovation production. Many of the economists have emphasized the increase of innovation activities at the end of the nineties. That's true, but some considerations have to be added due the implications of the results from the temporal segregation analysis. No sector has been predominant in terms of innovation activity in this period. Innovative ferment has been involved all the sectors in the, more or less, same size. It has been a general, little revolution common to all industrial sectors.

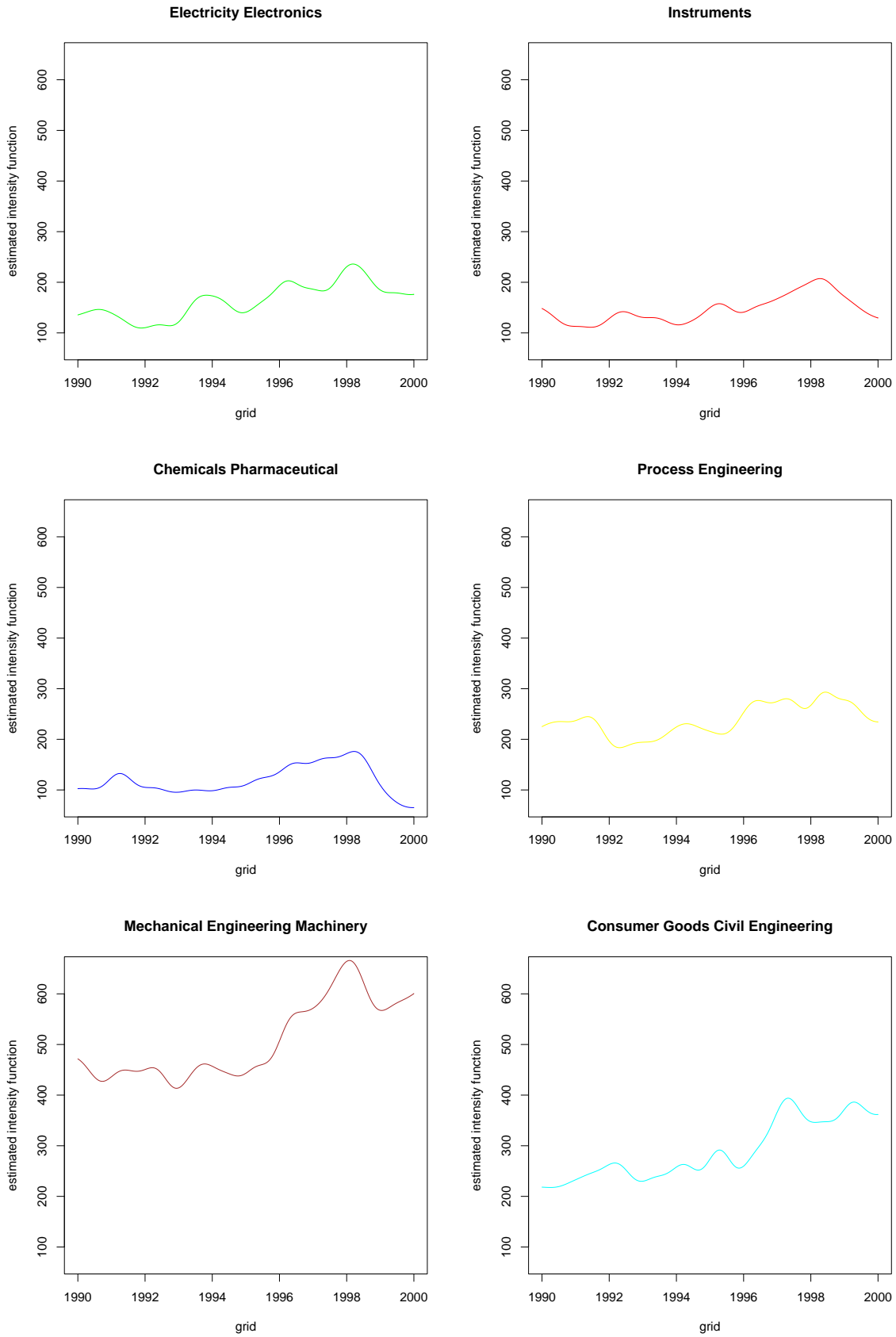


Figure 1.11: intensities

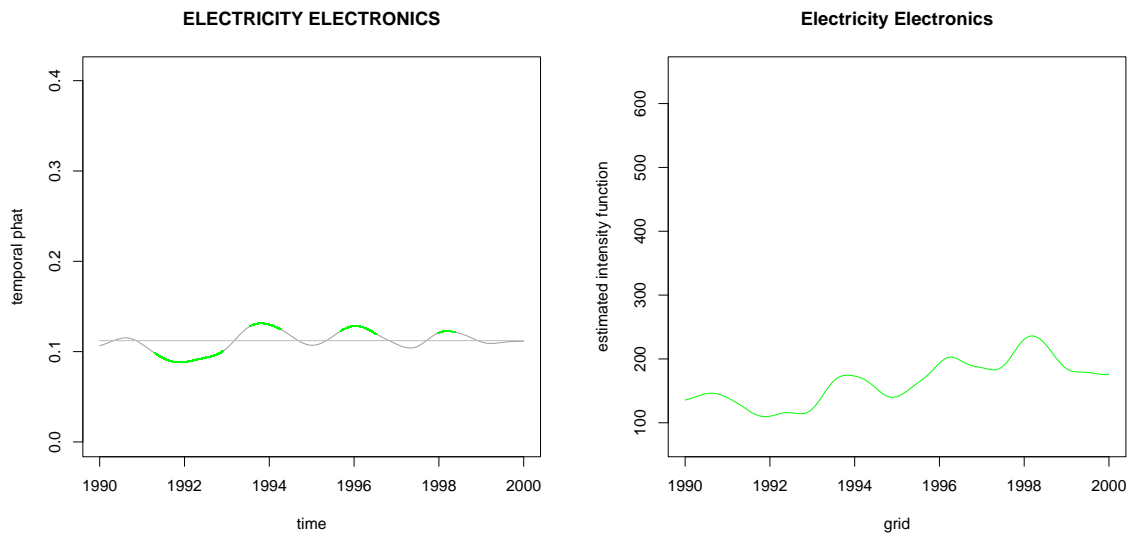


Figure 1.12: comparison1

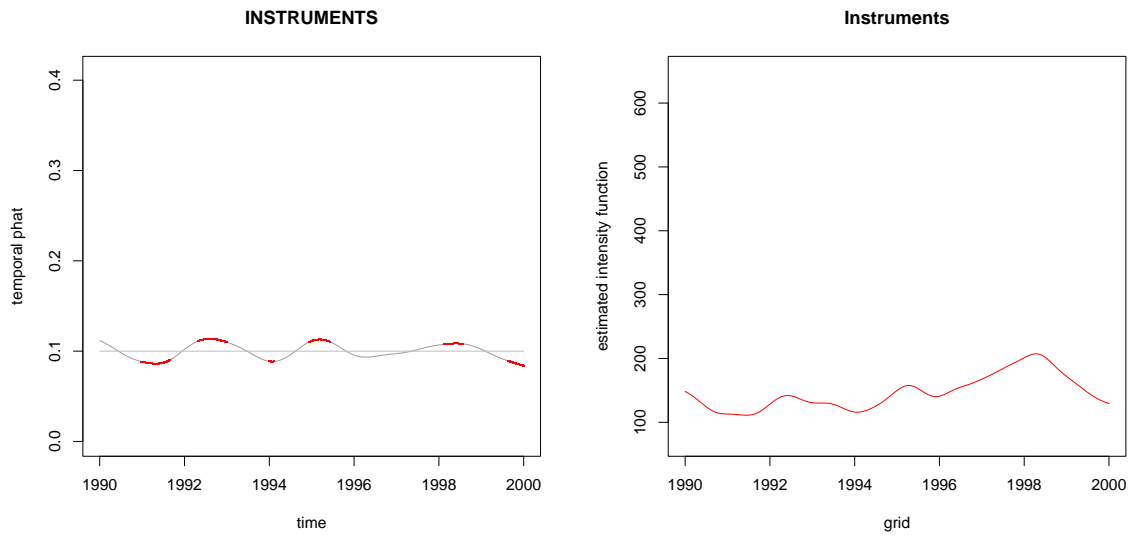


Figure 1.13: comparison2

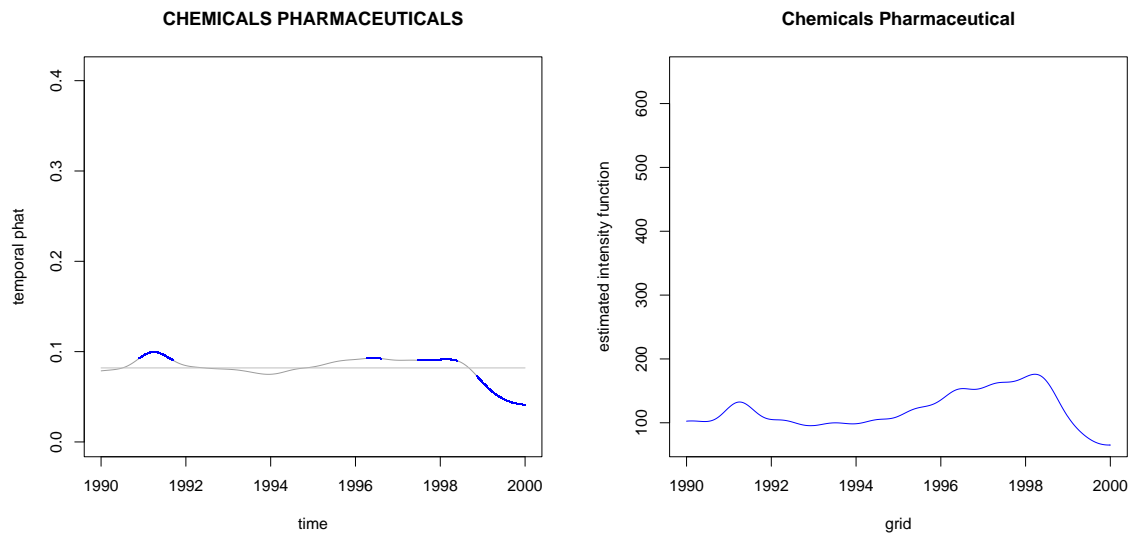


Figure 1.14: comparison3

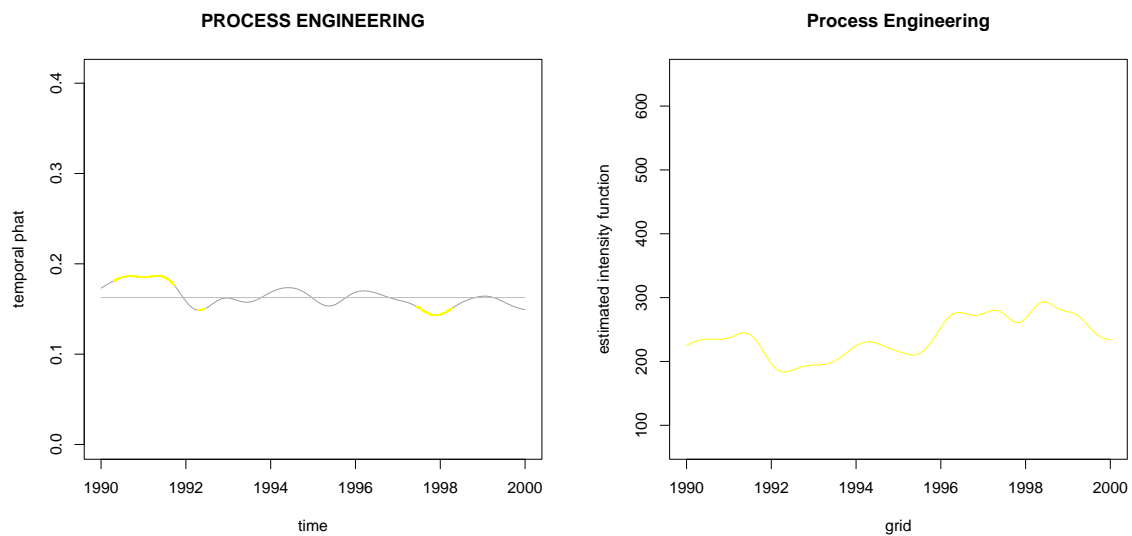


Figure 1.15: comparison4

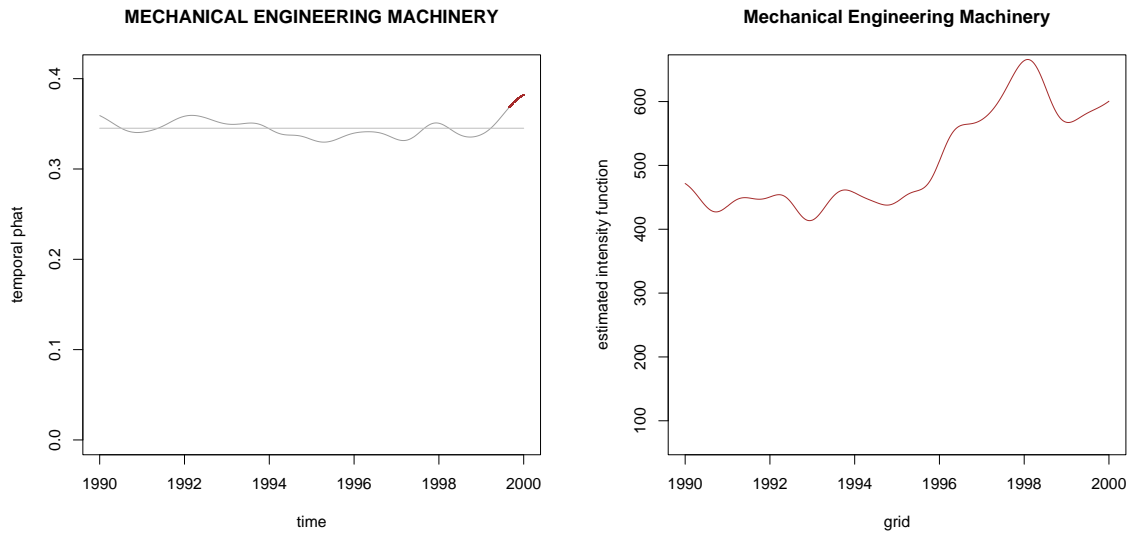


Figure 1.16: comparison5

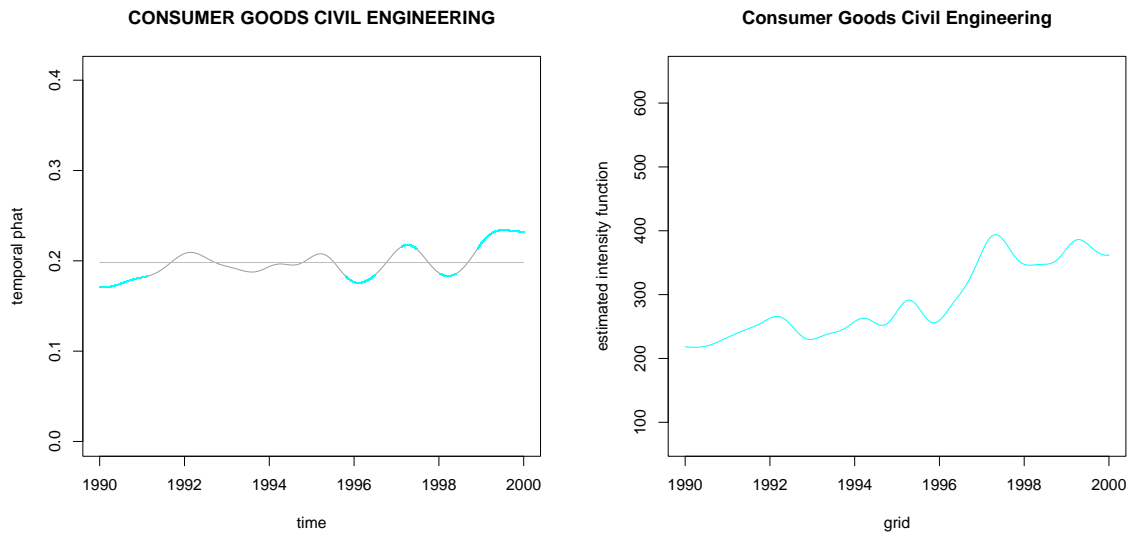


Figure 1.17: comparison6

Chapter 2

Semiparametric Intensity and Density Estimation

2.1 From Density to Intensity

So far we estimated the intensity function of the Poisson process non-parametrically with the kernel type estimator based on X_1, X_2, \dots, X_n observations:

$$\hat{\lambda}(x) = \sum_{i=1}^n h^{-1} K\left(\frac{X_i - x}{h}\right).$$

This procedure is very closed to the non-parametric kernel density estimation which provides the following estimator of the true density f :

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n h^{-1} K\left(\frac{X_i - x}{h}\right)$$

which differs from the first just by the normalizing factor n in the denominator.

Therefore the next step is to show and to develop a semiparametric procedure for density estimation and to extend the latter, in a second moment, to the intensity estimation in a very easy way as above; and finally, to apply it to the innovation dataset. The basic idea (Hjort and Glad, 1995,(18)) of the semiparametric approach is to combine an initial parametric density estimate with a kernel type estimate of the necessary correction factor. This method is designed to work better than the totally

non-parametric kernel estimator in a broad neighbourhood of a given parametric class of density, while not losing much in precision when the true density (or intensity in a second moment) is far from the parametric initial class.

The developed theory consists with to extend this method not only for the intensity estimation, but also for its use to analyze pooled data as in the innovation dataset. The parametric initial part is estimated using pooled data as considering it as the main common behaviour. The non-parametric kernel correction factor is, instead, estimated considering just one of the original sub-dataset and it becomes a measure of dissimilarity from the main behaviour.

2.2 Semiparametric Density Estimation

2.2.1 Introduction

Let X_1, X_2, \dots, X_n be independent observations from an unknown density f on the real line. The traditional nonparametric density estimator is

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n h^{-1} K \left(\frac{X_i - x}{h} \right) = \frac{1}{n} \sum_{i=1}^n K_h(X_i - x) \quad (2.2.1)$$

where $K_h(z) = h^{-1}K(h^{-1}z)$ and $K(z)$ is a kernel function, which is taken here to be a symmetric probability density with finite values of $\sigma_K^2 = \int z^2 K(z) dz$ and $R(K) = \int K(z)^2 dz$. The basic statistical properties are that (see Wand and Jones, 1995, (26)):

$$E\hat{f}(x) = f(x) + \frac{1}{2}\sigma_K^2 h^2 f''(x) + o(h^2) \quad (2.2.2)$$

$$Var\hat{f}(x) = R(K)(nh)^{-1}f(x) - \frac{f(x)^2}{n} + o((nh)^{-1}). \quad (2.2.3)$$

The integrated mean square error is of order $n^{-4/5}$ when h is proportional to $n^{-1/5}$. This procedure is totally nonparametric and impartial to any type of shapes of the underlying true density. The intention of the work of Hjort and Glad (1995,(18)) is to construct an alternative, but similar method, with same good properties, but better in

the broad vicinity of a given parametric family. As already written, they start out with a parametric density estimate $f(x, \hat{\theta})$ and then multiply with a nonparametric kernel type estimate of the correction factor $r(x) = \frac{f(x)}{f(x, \hat{\theta})}$. So their proposal is $\hat{r}(x) = \frac{n^{-1}K_h(X_i - x)}{f(X_i, \hat{\theta})}$, producing the final estimator

$$\hat{f}(x) = f(x, \hat{\theta})\hat{r}(x) = \frac{1}{n} \sum_{i=1}^n K_h(X_i - x) \frac{f(x, \hat{\theta})}{f(X_i, \hat{\theta})}. \quad (2.2.4)$$

It's not necessarily that the initial parametric estimate provides a serious approximation to the true density. In this case this method work well as the totally nonparametric kernel one. It works better if the parametric initial estimate is closed to the true density. The case of a constant start value for $f(x, \hat{\theta})$ corresponds to choosing a uniform distribution as the initial description, giving back the classic kernel estimator.

2.2.2 Nonparametric correction on a fixed start

Suppose f_0 is a fixed density, perhaps a crude guess of f . Write $f = f_0 r$. The idea is to estimate the nonparametric correction factor r via kernel smoothing as

$$\hat{r} = n^{-1} \sum_{i=1}^n K_h(X_i - x) / f_0(X_i)$$

which gives the estimator

$$\hat{f}(x) = f_0(x)\hat{r}(x) = \frac{1}{n} \sum_{i=1}^n K_h(X_i - x) \frac{f_0(x)}{f_0(X_i)}. \quad (2.2.5)$$

Note that a constant f_0 gives back the ordinary kernel estimator 2.2.1. The so-built estimator for the correction factor r has expected value

$$E[\hat{r}(x)] = \int K_h(y - x) f_0(y)^{-1} f(y) dy \quad (2.2.6)$$

$$= \int K(z) r(x + hz) dz = r(x) + \frac{1}{2} \sigma_K^2 h^2 r''(x) + O(h^4), \quad (2.2.7)$$

and variance

$$Var[\hat{r}(x)] = \frac{1}{n} \left[\int \frac{K_h(y-x)^2}{f_0(y)^2} f(y) dy - \{E\hat{r}(x)\}^2 \right] \quad (2.2.8)$$

$$= \frac{R(K)}{nh} \frac{f(x)}{f_0(x)^2} - \frac{r(x)^2}{n} + O(h/n) \quad (2.2.9)$$

and this shows that 2.2.5 has

$$bias \doteq \frac{1}{2} \sigma_K^2 h^2 r''(x) \quad \text{and} \quad variance \doteq \frac{R(K)}{nh} f(x) - \frac{f(x)^2}{n}.$$

Note that the variance is of the same size as that of the traditional non parametric estimator 2.2.1 to the order of approximation used, and the bias is of the same order h^2 , but proportional to $f_0 r''$ rather than f'' . The new estimator is better than the traditional one in all cases where $f_0 r''$ is smaller in size than $f'' = f_0'' r + 2f_0' r' + f_0 r''$. In cases where f_0 is already a good guess on f , expects r near constant and r'' small, this describes a certain neighbourhood of densities around f_0 where the new method is better than the traditional one.

2.2.3 Nonparametric correction on a parametric start

Let $f(x, \theta)$ be a given parametric family of densities then, the parametric start estimate is $f(x, \hat{\theta})$ where $\hat{\theta}$ is intended to be the maximum likelihood estimator. Even if a good choice for the initial start lead the estimator to perform better than the traditional nonparametrical kernel one, it works well also when f cannot be approximated by $f(x, \hat{\theta})$.

The task is to estimate the necessary correction function $f(x)/f(x, \hat{\theta})$ by kernel smoothing means and as suggested in the previous section $\hat{r}(x) = n^{-1} \sum_{i=1}^n K_h(X_i - x) / f(X_i, \hat{\theta})$.

In other words,

$$\hat{f}(x) = f(x, \hat{\theta}) \frac{\frac{1}{n} \sum_{i=1}^n K_h(X_i - x)}{f(X_i, \hat{\theta})}. \quad (2.2.10)$$

In order to understand to what extent the parametric estimation makes this estimator quantitatively different from the cleaner version 2.2.5, we bring in facts about the behaviour of the maximum likelihood estimator outside model conditions. It aims at a certain θ_0 , the least false value according to the Kullback-Leibler distance measure $\int f(x) \log f(x)/f(x, \theta) dx$ from true f to approximant $f(\cdot, \theta)$. Write $f_0(x) = f(x, \theta_0)$ for this best parametric approximant, and let $u_0(x) = \partial \log f(x, \theta_0)/\partial \theta$ be the score function evaluated at this parameter value. A Taylor expansion gives

$$\frac{f(x, \hat{\theta})}{f(X_i, \hat{\theta})} = \exp \left\{ \log f(x, \hat{\theta}) - \log f(X_i, \hat{\theta}) \right\} \quad (2.2.11)$$

$$\doteq \frac{f_0(x)}{f_0(X_i)} + \frac{f_0(x)}{f_0(X_i)} \{u_0(x) - u_0(X_i)\}' (\hat{\theta} - \theta_0), \quad (2.2.12)$$

leading to

$$\hat{f}(x) = \doteq \frac{1}{n} \sum_{i=1}^n K_h(X_i - x) \frac{f_0(x)}{f_0(X_i)} \left[1 - \{u_0(x) - u_0(X_i)\}' (\hat{\theta} - \theta_0) \right] \quad (2.2.13)$$

$$= f^*(x) + V_n(x). \quad (2.2.14)$$

Here f^* is as in 2.2.5, except for the fact that the f_0 function appearing here is not directly visible, and the $V_n(x)$ term stems from the parametric estimation variability. Representation 2.2.13, in concert with expressing $\hat{\theta} - \theta_0$ as an average of i.i.d. zero mean variables plus remainder term, can now be used to establish approximate bias and variance results for $\hat{f}(x)$.

Theorem 2.2.1. *Let $f_0(x) = f(x, \theta_0)$ be the best parametric approximant to f , and let $r = f/f_0$. The semiparametric estimator 2.2.10 has*

$$E[\hat{f}(x)] = f(x) + \frac{1}{2}\sigma_K^2 h^2 f_0(x)r''(x) + O(h^2/n + h^4 + n^{-2}) \quad (2.2.15)$$

$$Var[\hat{f}(x)] = R(K)(nh)^{-1}f(x) - \frac{f(x)^2}{n} + O(h/n + n^{-2}). \quad (2.2.16)$$

The detailed proof is in Hjort and Glad (1995, (18))

The result is remarkable in its simplicity; the sizes of bias and variance are only affected by parametric estimation noise to the quite small $O(h^2/n + n^{-2})$ order. The reason lies in expression 2.2.12. Not only is $\hat{\theta}$ close to θ_0 , but the $\hat{f}(x)$ estimator uses only X_i s that are close to x , making $u_0(X_i)$ close to $u_0(x)$.

Consistency of the density estimator requires both $h \rightarrow 0$ (forcing the bias towards zero) and $nh \rightarrow \infty$ (making the variance go to zero). The optimal size of h will later be seen to be proportional to $n^{-1/5}$. These observations match the traditional facts for the classic 2.2.1 estimator. Note also that if the parametric model happens to be accurate, then the r function is equal to 1, and the bias is only $O(h^4 + h^2/n)$.

Example: Normal Start Estimate

The normal start estimate is of the form $\hat{\sigma}^{-1}\phi(\hat{\sigma}^{-1}(x - \hat{\mu}))$, with ϕ the density of a Gaussian random variable, where one can use maximum likelihood estimates $\hat{\mu} = n^{-1}\sum_{i=1}^n X_i$ and $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n (X_i - \hat{\mu})^2$ (or the de-biased version with denominator $n - 1$). In view of generality of the proposition above quite general estimators are allowed, without changing the basic structure of bias and variance of $\hat{f}(x)$. One might for example wish to use robust estimates of mean and standard deviation. In any case the new semiparametric density estimator is

$$\hat{f}(x) = \frac{1}{\hat{\sigma}} \phi\left(\frac{x - \hat{\mu}}{\hat{\sigma}}\right) \frac{\frac{1}{n} \sum_{i=1}^n K_h(X_i - x)}{\frac{1}{\hat{\sigma}} \phi\left(\frac{X_i - \hat{\mu}}{\hat{\sigma}}\right)} \quad (2.2.17)$$

$$= \frac{1}{n} \sum_{i=1}^n K_h(X_i - x) \frac{\exp\left\{-\frac{1}{2}(x - \hat{\mu})^2 / \hat{\sigma}^2\right\}}{\exp\left\{-\frac{1}{2}(X_i - \hat{\mu})^2 / \hat{\sigma}^2\right\}} \quad (2.2.18)$$

Note that its implementation is straightforward.

Example: Log-Normal Start Estimate

One option for positive data is to start with a log-normal approximation and then multiply with a correction factor. The result is

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K_h(X_i - x) \frac{\exp\left\{-\frac{1}{2}(x - \hat{\mu})^2 / \hat{\sigma}^2\right\} X_i}{\exp\left\{-\frac{1}{2}(X_i - \hat{\mu})^2 / \hat{\sigma}^2\right\} x}. \quad (2.2.19)$$

$$(2.2.20)$$

Example: Gamma Start Estimate

A version of the general method which should work well for positive data from perhaps unimodal and right-skewed distribution is to start with a gamma distribution approximation. The final estimator is then of the form

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n K_h(X_i - x) \left(\frac{x}{X_i}\right)^{\hat{\alpha}-1} \exp\left\{-\hat{\beta}(x - X_i)\right\}, \quad (2.2.21)$$

for example with moment estimates for the gamma parameters.

Remark

This method can be used for any given parametric model. It is intuitively clear that the method works the best in cases where the model employed is not too far from covering the truth. One could think of ways of automatising the choice of the parametric vehicle model, through suitable goodness of fit measures, thereby obtaining an overall adaptive density estimator, but this aim is not pursued here.

2.2.4 Comparison with the traditional kernel density estimator

In this section the performance of the new estimator is compared to that of the usual 2.2.1 estimator.

Expressions can be found for the leading terms of the integrated mean squared errors of the usual kernel estimator 2.2.1 and the new estimator 2.2.10, using respectively 2.2.2, 2.2.3 and theorem 2.2.1. So the results for the AMISE (approximate mean integrated square error) are:

$$\text{amise for } \hat{f}_{trad} = \frac{1}{4}\sigma_K^4 h^4 R_{trad}(f) + R(K)(nh)^{-1} \quad (2.2.22)$$

$$\text{amise for } \hat{f}_{new} = \frac{1}{4}\sigma_K^4 h^4 R_{new}(f) + R(K)(nh)^{-1} \quad (2.2.23)$$

featuring ‘roughness’ functionals as

$$R_{trad}(f) = \int \{f''(x)\}^2 dx \quad (2.2.24)$$

$$R_{new}(f) = \int \{f_0(x)r''(x)\}^2 dx. \quad (2.2.25)$$

The new estimator is better, in the sense of approximate (leading terms) integrated

mean squared error, whenever $R_{new}(f)$ is smaller than $R_{trad}(f)$. This defines a non-parametric neighbourhood of densities around the parametric class. When f belongs to this neighbourhood, \hat{f}_{new} is better than \hat{f}_{trad} when the same K and the same h are used in the two estimators. In such a case the new estimator can be made even better by choosing an appropriate h as suggested later.

2.2.5 Choosing Smoothing Parameter

This method is defined in terms of a kernel function K and a bandwidth or smoothing parameter h . Choosing h is the more crucial problem, and methods for doing this parallel, but by necessity become harder than the well-developed ones for the traditional 2.2.1 estimator which is the special case of a constant initial estimator.

Minimising Estimated AMISE

A useful idea related to the previous calculations is to estimate the approximate MISE of 2.2.23 directly, that is, producing the curve

$$\widehat{amise}(h) = bcv(h) = \frac{1}{4}\sigma_K^4 h^4 \left\{ R_{new}(h) - \frac{R(K'')}{nh^5} \right\} + \frac{R(K)}{nh} \quad (2.2.26)$$

including for emphasis h in the notation for the roughness estimate. This function must now be computed for a range of h -values, up to some upper limit h_{os} , the ‘over-smoothing’ bandwidth. Usually this strategy is called *biased cross validation*, although nothing seems to be cross validated per-se. The *bcv* name derives rather from formula-wise similarity to unbiased cross validation, see below, and the desire to estimate the biased approximation AMISE to the true MISE.

Nearly Unbiased Cross Validation

A popular technique for the traditional kernel estimator is that of unbiased least squares cross validation, minimising an unbiased estimate of the exact MISE as a function of bandwidth. A version of this idea can be carried out for the new estimator as well. The crux is to estimate

$$\text{mise}(h) - R(f) = E \left\{ \int \hat{f}^2 dx - 2 \int f \hat{f} dx \right\} \quad (2.2.27)$$

with

$$\text{ucv}(h) = \int \hat{f}_h(x)^2 dx - \frac{2}{n} \sum_{i=1}^n \hat{f}_{h,(i)}(X_i). \quad (2.2.28)$$

Here h is included in the notation for clarity, and $\hat{f}_{h,(i)}$ is the estimator constructed from the diminished data set that excludes X_i . The function to compute is

$$\begin{aligned} \text{ucv}(h) = & \frac{1}{n^2} \sum_{i,j} \frac{1}{f(X_i, \hat{\theta}) f(X_j, \hat{\theta})} \int f(x, \hat{\theta})^2 K_h(x - X_i) K_h(x - X_j) dx \\ & - \frac{2}{n(n-1)} \sum_{i,j} K_h(X_i - X_j) \frac{f(X_i, \hat{\theta}_{(i)})}{f(X_j, \hat{\theta}_{(j)})} \end{aligned} \quad (2.2.29)$$

where $\hat{\theta}_{(i)}$ is computed without X_i . In the case of the normal start method 2.2.17 with normal kernel $K = \phi$ a formula for the first term here is given below.

It turns out that $\text{ucv}(h)$ is nearly, but not exactly unbiased for $\text{mise}(h) - R(f)$. We have

$$E \int f \hat{f} dx = E \int f(x) K_h(X_1 - x) \frac{f(x, \hat{\theta})}{f(X_1, \hat{\theta})} dx \quad (2.2.30)$$

$$= E \int \int f(x) f(y) K_h(y - x) \frac{f(x, \hat{\theta}(y, X_2, \dots, X_n))}{f(y, \hat{\theta}(y, X_2, \dots, X_n))} dx dy, \quad (2.2.31)$$

which is subtly different from

$$\begin{aligned} E \frac{1}{n} \sum_{i=1}^n \hat{f}_{(i)}(X_i) &= EK_h(X_2 - X_1) \frac{f(X_1, \hat{\theta}(y, X_2, \dots, X_n))}{f(X_1, \hat{\theta}(X_2, X_2, \dots, X_n))} & (2.2.32) \\ &= E \int \int f(x) f(y) K_h(y - x) \frac{f(x, \hat{\theta}(y, X_3, \dots, X_n))}{f(y, \hat{\theta}(y, X_3, \dots, X_n))} dx dy & (2.2.33) \end{aligned}$$

The difference is minuscule, however, and choosing h to minimise the $ucv(h)$ function, among $h \leq h_{os}$ for a suitable over-smoothing upper limit, remains a useful non-parametric option.

2.2.6 Accuracy of the estimated correction factor

This machinery can be used for model exploration purposes, by inspecting the correction factor against x for various potential models. A model's adequacy could be inspected by looking at a plot of $\hat{r}(x)$, perhaps with a pointwise confidence band, to see if $r(x) = 1$ is reasonable.

It is also informative to plot the log-correction factor $\log \hat{r}(x)$, to see how far from zero it is.

A nice graphical goodness of fit method emerges from the previous results by plotting

$$Z(x) = \frac{\log \hat{r}(x) + \frac{1}{2} R(K)(nh)^{-1} f(x, \hat{\theta})^{-1}}{\left\{ R(K)(nh)^{-1} f(x, \hat{\theta})^{-1} \right\}^{1/2}} \quad (2.2.34)$$

against x , possibly with a more accurate denominator. Under model conditions this should be approximately distributed as a standard normal for each x , that is, the $Z(x)$ curve should stay within ± 1.96 about 95% of the time.

2.3 Semiparametric Intensity Estimation

2.3.1 Introduction

The non-parametric estimator for the intensity function of the Poisson process proposed in the previous section uses a kernel type estimator based on X_1, X_2, \dots, X_n independent observations:

$$\hat{\lambda}(x) = \sum_{i=1}^n h^{-1} K\left(\frac{X_i - x}{h}\right)$$

where K is the Kernel, a symmetric density function, and h the smoothing parameter. The similarity with the kernel non-parametric density estimation is evident. The only difference is that the normalizing factor n is present in the density estimator. Starting from this similarity I extended the semiparametric procedure from density to intensity estimation. In a very simple way as we will see.

The basic idea is the same as for the semiparametric density estimation: view the intensity as a combination of a parametric function times a correction factor,

$$\lambda(x) = g(x)r(x)$$

which leads to combine an initial parametric start maximum likelihood estimate with a correction kernel-type estimated factor.

2.3.2 A parametric function corrected by a kernel factor

The intensity function of a Poisson point process can be modeled with a parametric function $\lambda(x) = g(x, \theta)$. A maximum likelihood approach is possible: here, θ could be estimated maximising the likelihood of the Poisson point process (2.3.1) with intensity function $\lambda(x) = g(x, \theta)$.

An instance in which the likelihood function is tractable is the inhomogeneous Poisson process with intensity function $\lambda(x)$. Essentially because the distribution associated

with a partial realization $X = X_1, \dots, X_n$, of this process on a finite region A can be factorized as the product of a Poisson distribution with mean $\mu = \int_A \lambda(x)dx$ for the number of events n , and a set of mutually independent locations x_i whose common distribution has density $\lambda(x)/\mu$. Hence, the log-likelihood for $\lambda(\cdot)$ based on data X is

$$ll(\lambda) = \sum_{i=1}^n \log \lambda(x_i) - \int_A \lambda(x)dx \quad (2.3.1)$$

The maximising process could be carried out numerically using the ‘Nelder-Mead’ algorithm.

I want now to introduce a correction factor in order to improve the parametric estimation procedure. I would like the correction factor supplies a better estimate if and just where the parametric estimates doesn’t mach the true intensity; but it must be inactive if my start model is the correct one! The idea is very closed to Hjort and Glad (1995,(18)). The proposed semiparametric estimator is:

$$\hat{\lambda}(x) = \hat{g}(x)\hat{r}(x) \quad (2.3.2)$$

$$= g(x, \hat{\theta}) \sum_{i=1}^n \frac{K_h(X_i - x)}{g(X_i, \hat{\theta})} \quad (2.3.3)$$

where $\hat{g}(x)$ is the initial parametric model and $\hat{\theta}$ is obtained maximising the likelihood of the Poisson point process (2.3.1) with intensity function $\lambda(x) = g(x, \theta)$. Afterwards that, the correction factor is applied using a kernel type estimator. It will drive the parametric model when it can’t supply a good estimate for the intensity of the Poisson point process. Crucial, as in every non-parametric kernel setting, is the choice of the bandwidth.

2.3.3 Choosing the Optimal Bandwidth

The Mean Integrated Square Error (MISE) plays a very important role in non-parametric kernel estimation. It is a valid measure of the goodness of our estimator. So, a popular technique is to use this error measure to drive the choice of the bandwidth: we will choose the value of h which minimize the MISE, better a nearly unbiased estimate of it. This procedure is called *nearly unbiased cross validation* and it is used in density estimation and, with a very little variation, in intensity estimation. So the next step is to minimize an unbiased estimate of the exact MISE as a function of bandwidth

$$ucv(h) = \int \hat{\lambda}_h(x)^2 dx - 2 \sum_{i=1}^n \hat{\lambda}_{h,(i)}(X_i). \quad (2.3.4)$$

where $\hat{\lambda}_{h,(i)}$ is the estimator constructed from the diminished data set that excludes X_i .

The role of the value of h become more interesting here: if the *ucv* procedure selects a small value h , it means that the correction factor must be active very often let the parametric start be inappropriate. Instead, if a very large value for h selected leads to a very smooth behaviour of the correction factor, the parametric start chosen is quite good!

2.3.4 Simulation Study

The new semiparametric intensity estimator is tested here in three different situations. The value of the optimal bandwidth h and the behaviour of the correction factor will help us to understand if the parametric start was good or not.

I simulated data from a temporal Poisson point process with intensity function

$$\begin{aligned} \log \lambda(x)_{true} &= \theta_1 + \theta_1 x + \theta_3 \cos(wx) + \theta_4 \sin(wx) \\ &+ \theta_5 \cos(2wx) + \theta_6 \sin(2wx) + \theta_7 \cos(3wx) + \theta_8 \sin(3wx); \end{aligned} \quad (2.3.5)$$

where $w = 2\pi/365$.

In a first simulation study I use the exact parametric start function, so my semiparametric estimator overwrites exactly the parametric one and the correction factor is, consequently, always equal to 1.

In a second simulation study I use a wrong parametric start, so the semiparametric estimator must be active and the correction factor shows where and how the parametric form is not correct.

Finally in the third simulation study the parametric start used is wrong and the correction factor shows that it misses a linear trend in time.

Model 1

Considering a temporal Poisson point process with intensity function as in 2.3.5 with true parameters reported in table 2.1 in the row called ‘TRUE’ and plotted in fig. 2.1 in black. I simulated points from such process and I use them to estimate the intensity parametrically making the hypothesis that

$$\begin{aligned} \lambda(x) = g(x, \theta) = & \exp[\theta_1 + \theta_1 x + \theta_3 \cos(wx) + \theta_4 \sin(wx) \\ & + \theta_5 \cos(2wx) + \theta_6 \sin(2wx) + \theta_7 \cos(3wx) + \theta_8 \sin(3wx)] \end{aligned} \quad (2.3.6)$$

which correspond to make the right choice. The estimated θ s are obtained maximising the likelihood of the Poisson point process (2.3.1). Maximum likelihood estimates are reported in the last row of table 2.1. The ML estimated intensity function is plotted in fig. 2.1. A full kernel non-parametric estimated intensity is plotted in green. Now I use the new semiparametric estimator. The first step is to decide from which parametric model we have to start the procedure. We use the same, correct, parametric form 2.3.6. So we know that our hypothesis is the true one. The semiparametric estimator

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8
TRUE	0.1	0.0009	-0.055	-0.046	-0.113	0.094	0.13	0.095
ML	0.1269	0.0008	-0.0535	-0.0224	-0.1102	0.1549	0.1595	0.0960

Table 2.1: model 1

becomes

$$\hat{\lambda}(x) = g(x, \hat{\theta}) \sum_{i=1}^n \frac{K_h(X_i - x)}{g(X_i, \hat{\theta})}$$

with

(2.3.7)

$$g(x, \theta) = \exp[\theta_1 + \theta_1 x + \theta_3 \cos(wx) + \theta_4 \sin(wx) + \theta_5 \cos(2wx) + \theta_6 \sin(2wx) + \theta_7 \cos(3wx) + \theta_8 \sin(3wx)]$$
(2.3.8)

and it uses the ML estimates for θ already obtained. I need now to find the optimal bandwidth using the *unbiased cross validation* procedure which consists in minimizing the equation 2.3.4. The chosen h is equal to 248 time periods which is a very big value that suggests that the correction factor didn't work so well due to the good parametric start. Now I am able to plot the semiparametrically estimated intensity function (in fig. 2.1 in blue). We note that it overwrites the parametric red one: the starting model was already very good, the correction factor didn't work and the semiparametric estimator worked well as the parametric one since the model hypothesis was correct. For this reason the correction factor plotted in cyan is always equal to 1.

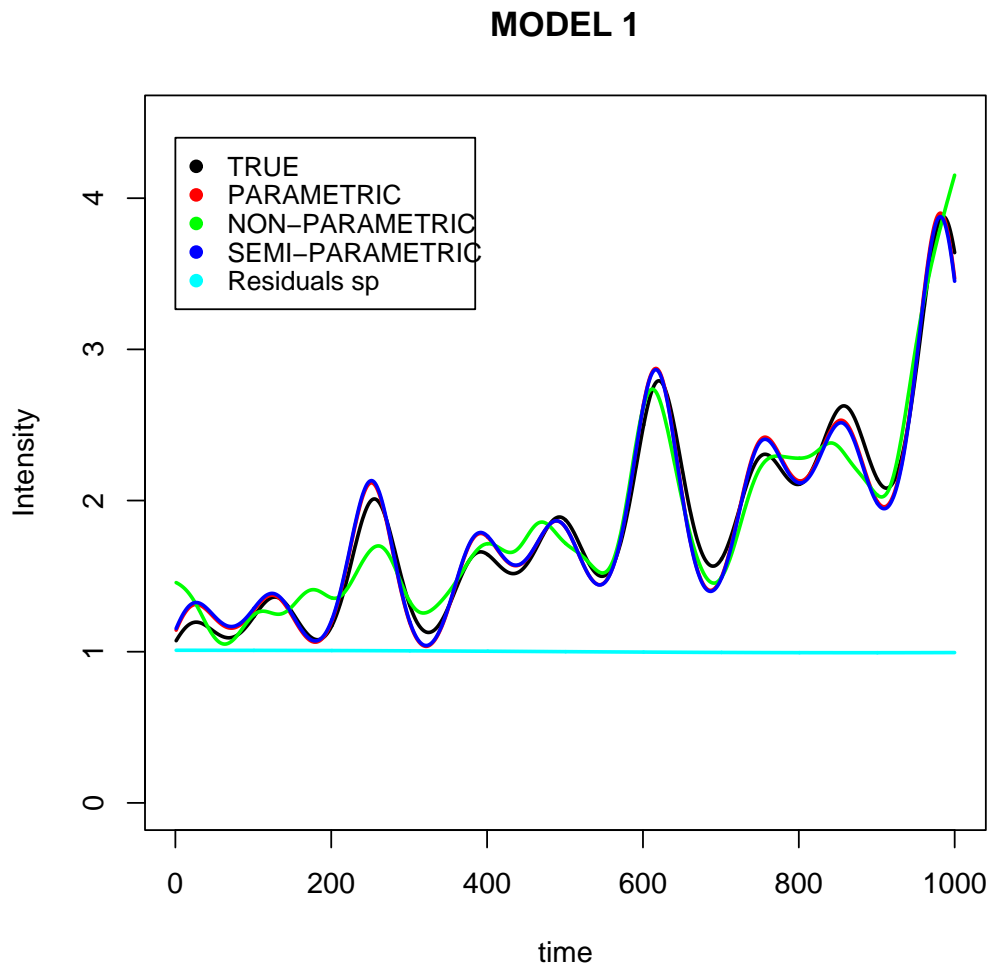


Figure 2.1: Model 1. (*Residuals sp* are the residuals of the semiparametric estimation)

Model 2

Now I test the behaviour of the new semiparametric estimator starting from a wrong parametric hypothesis. Data are simulated from the same Poisson point process as above with same intensity function and same parameters. Intensity plotted in black in Fig.2.2

The parametric inference was carried out under the following wrong hypothesis:

$$\begin{aligned} \lambda(x) = g(x, \theta) = & \exp[\theta_1 + \theta_1 x + \theta_3 \cos(wx) + \theta_4 \sin(wx) \\ & + \theta_5 \cos(2wx) + \theta_6 \sin(2wx)] \end{aligned} \quad (2.3.9)$$

which doesn't capture some stagionalities present in the true model. Maximum likelihood inference provides estimates in Table 2.2 and the parametrically estimated intensity function is plotted in red in Fig. 2.2. A full kernel non-parametric estimated intensity is plotted in green.

TRUE	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8
	0.1	0.0009	-0.055	-0.046	-0.113	0.094	0.13	0.095
ML	0.0679	0.0009	-0.0132	-0.0647	-0.0943	0.0919281024	XXX	XXX

Table 2.2: model 2

I expect the semiparametric estimator recognizes that the parametric start is not precise and it moves toward the non-parametric estimates giving more weight to the non-parametric correction factor. My guess is reinforced by the optimal bandwidth selected: 33 which is a small value which allows the correction factor be active and suggesting that the parametric part fits the data badly. The semi-parametrically estimated intensity function plotted in blue in Fig. 2.2 validates the expected ideas. And the plot of the correction factor suggests where it must work more and how and where we can improve now the parametric starting model.

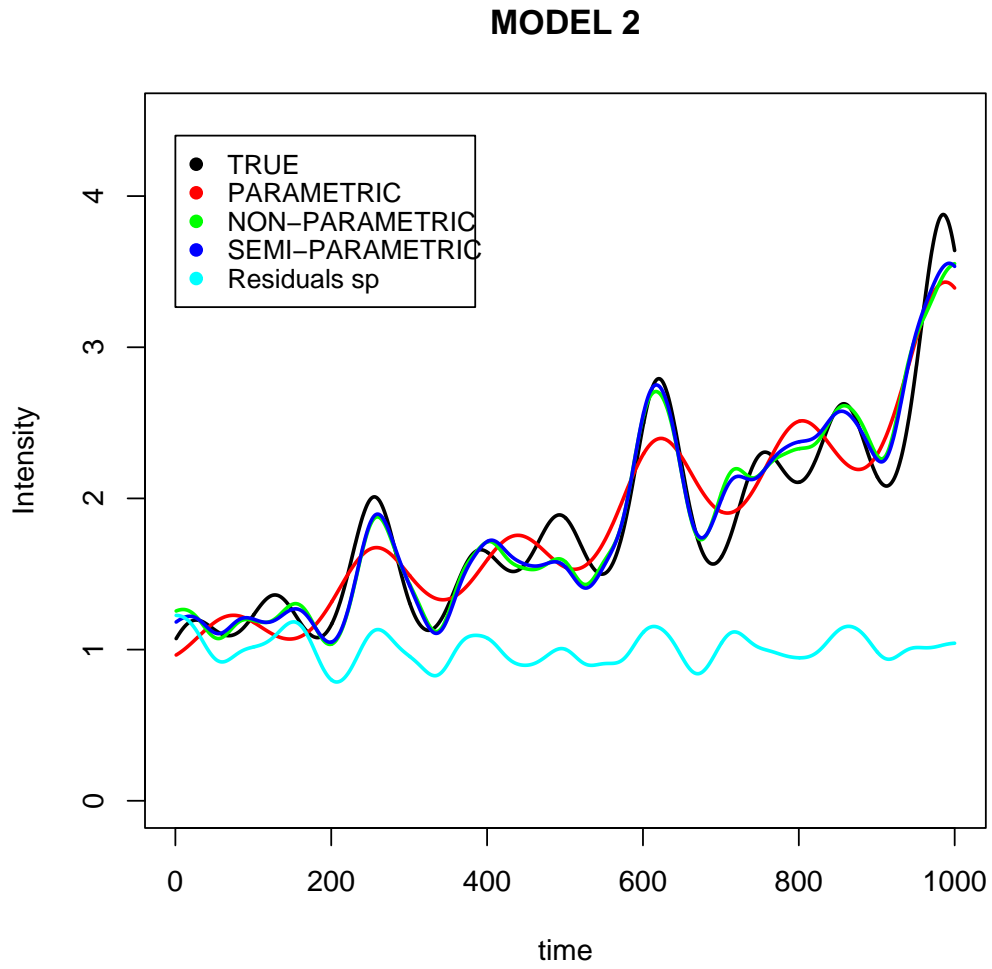


Figure 2.2: Model 2

Model 3

In this last simulation, likewise model 2, a wrong parametric model is used, dropping from the true model the linear temporal trend. The parametric model assumed is now:

$$\begin{aligned} \lambda(x) = g(x, \theta) = & \exp[\theta_1 + \theta_3 \cos(wx) + \theta_4 \sin(wx) \\ & + \theta_5 \cos(2wx) + \theta_6 \sin(2wx) + \theta_7 \cos(3wx) + \theta_8 \sin(3wx)] \end{aligned} \quad (2.3.10)$$

whose parameters were maximum likelihood estimated and reported in table 2.3. The red plotted parametric estimated intensity in Fig. 2.3 shows explicitly the lack of the linear trend in the modeling. As usual a full kernel non-parametric estimated intensity is plotted in green.

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8
TRUE	0.1	0.0009	-0.055	-0.046	-0.113	0.094	0.13	0.095
ML	0.5762	XXX	-0.1042	-0.1064	-0.1201	0.0691	0.2124	0.0630

Table 2.3: model 3

What we expect now from the correction factor is to introduce just temporal trend allowing the good performance of the parametric start in modeling the stagionalities. And it works in this direction as we can see in Fig. 2.3 where the semiparametric estimator performed better than the full non parametric one in allowing the curvatures, and better than the parametric one in considering the temporal linear trend. The new semiparametric estimator captures the best behaviours of each of the two other types of estimators. The plot of the correction factor suggests quite explicitly what it is wrong and how we can improve the parametric model.

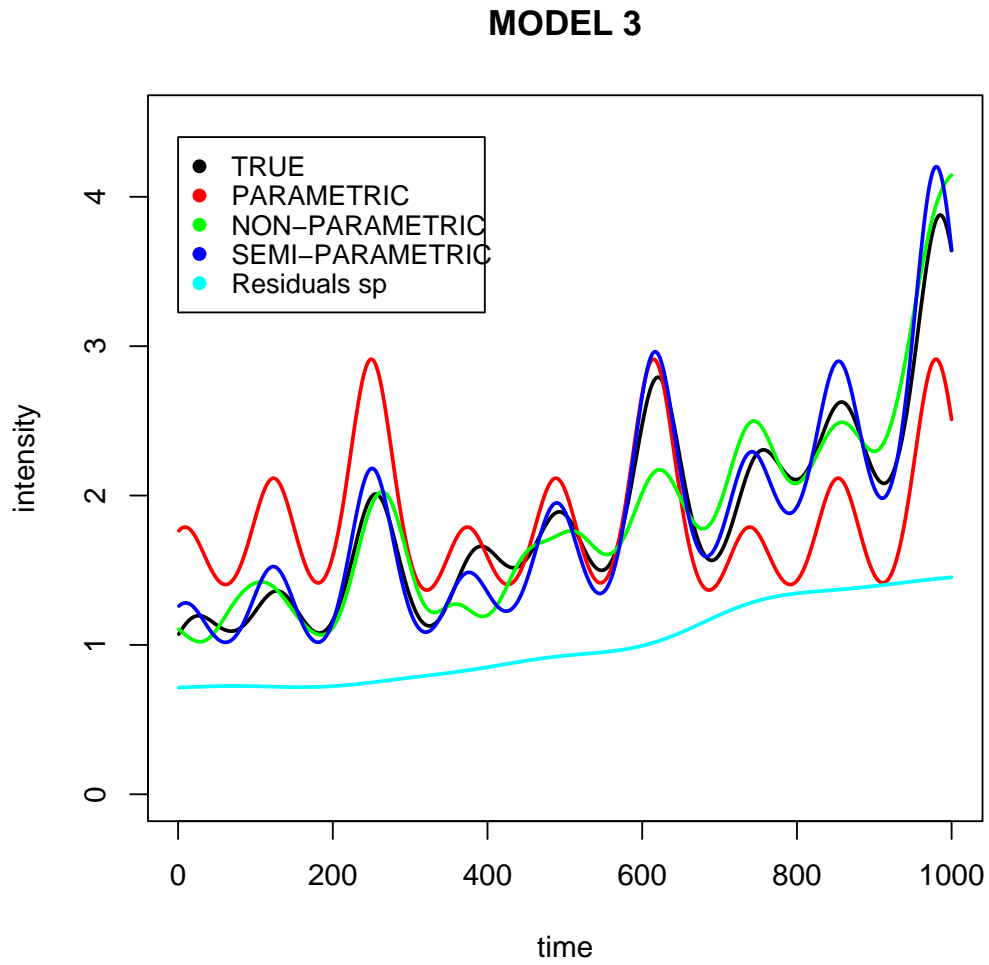


Figure 2.3: Model 3

2.3.5 Conclusions

For each of the previous models I simulated 100 datasets and for each of them I applied the three estimators and computed their performances in terms of Integrated Square Errors (ISE)

$$ISE = \int \lambda(x) - \hat{\lambda}(x) dx.$$

For each of the three scenarios the performances of semiparametric estimator are statistically significant bigger then the others.

2.4 Semiparametric Intensities Estimation using *pooled* data

2.4.1 Introduction

Innovation dataset is a typical example of *pooled* data: the complete dataset is composed of six sub-datasets, one for each industrial sector, in other words, one for each of the component process of the multivariate point process.

The idea is to extend the semiparametric intensity estimator in this setting too. I think that the intensities of the components of the multivariate point process share a common behaviour but each of them have a specific feature too. This leads us to consider the parametric part of the estimator as a measure of such common pattern and the non-parametric part as a description of the specific sector behaviour. So, the parametric part will use the whole dataset, the pooled data, in the estimation phase, instead the non-parametric part will consider just data for the specific sector.

2.4.2 *Pooled* data: the semiparametric approach

Let $\lambda_k(x)$ $k = 1, \dots, m$ be the intensities of the m components of the multivariate inhomogeneous Poisson point processes:

$$\begin{aligned}\hat{\lambda}_k(x) &= g(x, \hat{\theta}) \hat{r}_k(x) \\ &= g(x, \hat{\theta}) \sum_{i=1}^{n_k} \frac{K_{h_k}(X_i - x)}{g(X_i, \hat{\theta})}\end{aligned}\quad (2.4.1)$$

Here I want the parametric part to capture the common behaviour, so the estimated $\hat{\theta}$ is obtained maximising the log-likelihood (2.4.2) of the whole Poisson point process with intensity function $\lambda = g(x, \theta)$ and using the *pooled* data: the whole dataset;

$$ll(\lambda) = \sum_{i=1}^n \log \lambda(x_i) - \int_A \lambda(x) dx \quad (2.4.2)$$

The maximising process is carried out numerically using the ‘Nelder-Mead’ algorithm and can use a large number of observations making the estimation efficient, but just for the common pattern! Then I need to model the pattern that make the component processes different from each others. This specific behaviour is captured by the correction factor and to estimate it, I use just the data that come from the considered component and the choice of the optimal bandwidth is obtained minimising the *ucv* function, but just using the specific subdataset as showed in 2.4.3

$$ucv(h_k) = \int \hat{\lambda}_{k,h_k}(x)^2 dx - 2 \sum_{i=1}^{n_k} \hat{\lambda}_{h_k,(i)}(X_i). \quad (2.4.3)$$

where $\hat{\lambda}_{h_k,(i)}$ is the estimator constructed from the diminished data set that excludes X_i .

The logic hence is changed. In the previous section the correction factor $r(x)$ could supply better estimates when the parametric part is not appropriate. Here the correction factor $r(x)$ could capture departures of each component process from the common behaviour estimated by the parametric part using the pooled dataset.

2.4.3 Simulation Study

I simulated data from a multivariate inhomogeneous Poisson process with 6 component processes. The main and common pattern is modelled by this intensity:

$$\begin{aligned} \log \lambda(x) &= \theta_1 + \theta_1 x + \theta_3 \cos(wx) + \theta_4 \sin(wx) \\ &+ \theta_5 \cos(2wx) + \theta_6 \sin(2wx) + \theta_7 \cos(3wx) + \theta_8 \sin(3wx) \end{aligned} \quad (2.4.4)$$

$w = \frac{2\pi}{365}$. Then each of the six intensities from which I simulated data are modelled as:

$$\log \lambda_k(x) = \log \lambda(x) + \beta_{1,k} + \beta_{2,k}x + \beta_{3,k}x^2 \quad (2.4.5)$$

so each intensity λ_k is different from on others while sharing a common pattern designed by 2.4.4. The resulting intensities are plotted in fig. 2.4 where the black line is the common behaviour called *family*.

The first step consists in estimating θ , a vector of parameters of the parametric start, maximising the log-likelihood (2.4.2) of the whole Poisson point process with intensity function $\lambda = g(x, \theta)$ and using the *pooled* data: the whole dataset. Then I have to choose the optimal bandwidth with the *ucv* criterium (2.4.3) obtaining six optimal bandwidths, each for any component. Then I am able to use the correction factor and plot the estimated intensities. Just as an example I show the procedure for intensity 1 and intensity 6.

Intensity 1

In fig. 2.5 the common pattern is plotted in black and the true intensity of the first process in red .

Parametric Estimation: I make the hypothesis (partially wrong) the intensity 1 could

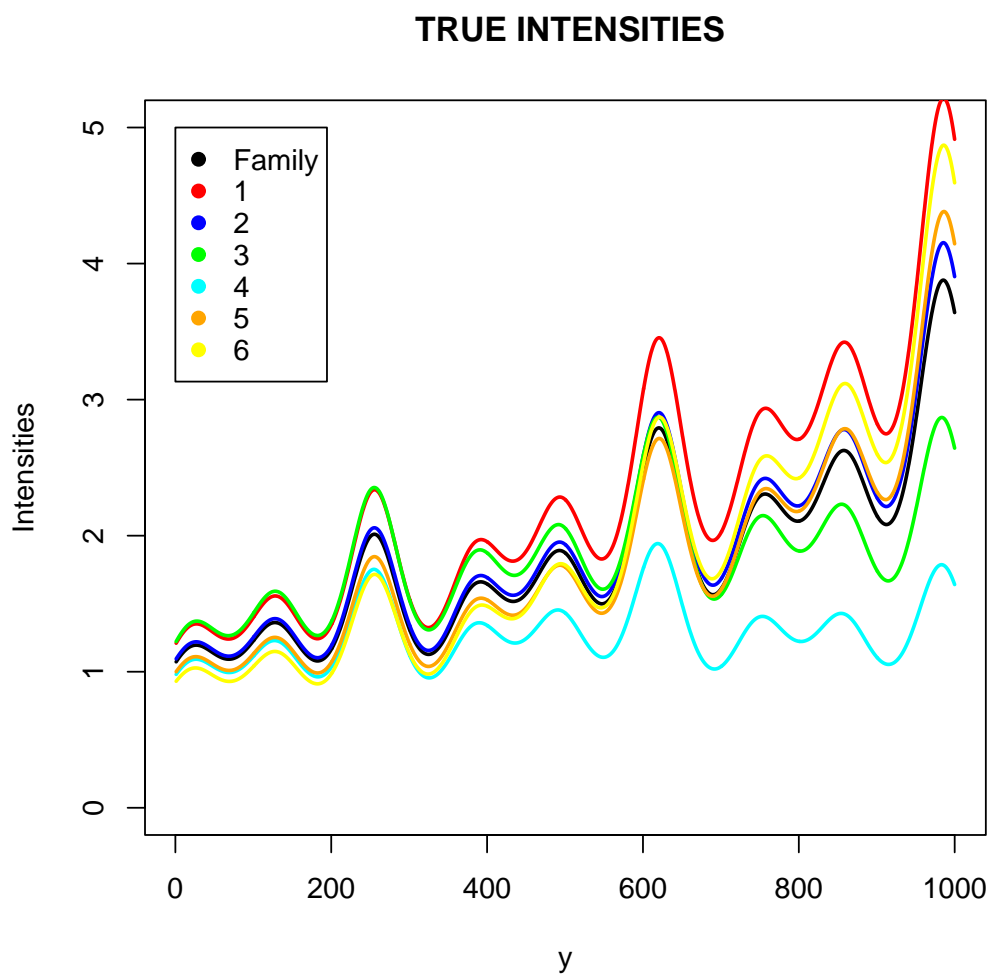


Figure 2.4: True Intensities

be modelled parametrically as:

$$\begin{aligned} \log \lambda_1(x) &= \theta_1 + \theta_1 x + \theta_3 \cos(wx) + \theta_4 \sin(wx) \\ &+ \theta_5 \cos(2wx) + \theta_6 \sin(2wx) + \theta_7 \cos(3wx) + \theta_8 \sin(3wx) \end{aligned} \quad (2.4.6)$$

which doesn't consider the specific variations. I optimize the log-likelihood in order to obtain θ estimated using just the data simulated from the process 1. This traditional parametrically estimated intensity is plotted in green.

Applying the semiparametric estimator in 2.4.1 we make the hypothesis that

$$\begin{aligned} \log g(x, \theta) &= \theta_1 + \theta_1 x + \theta_3 \cos(wx) + \theta_4 \sin(wx) \\ &+ \theta_5 \cos(2wx) + \theta_6 \sin(2wx) + \theta_7 \cos(3wx) + \theta_8 \sin(3wx) \end{aligned} \quad (2.4.7)$$

which wants to model just the shared features and now θ are obtained as before but now using the pooled dataset: data which come from all component processes. This parametric start will be multiply by the non-parametric kernel correction factor which uses just data from process 1. The semiparametrically estimated intensity is plotted in blue and it is very closed to the true density.

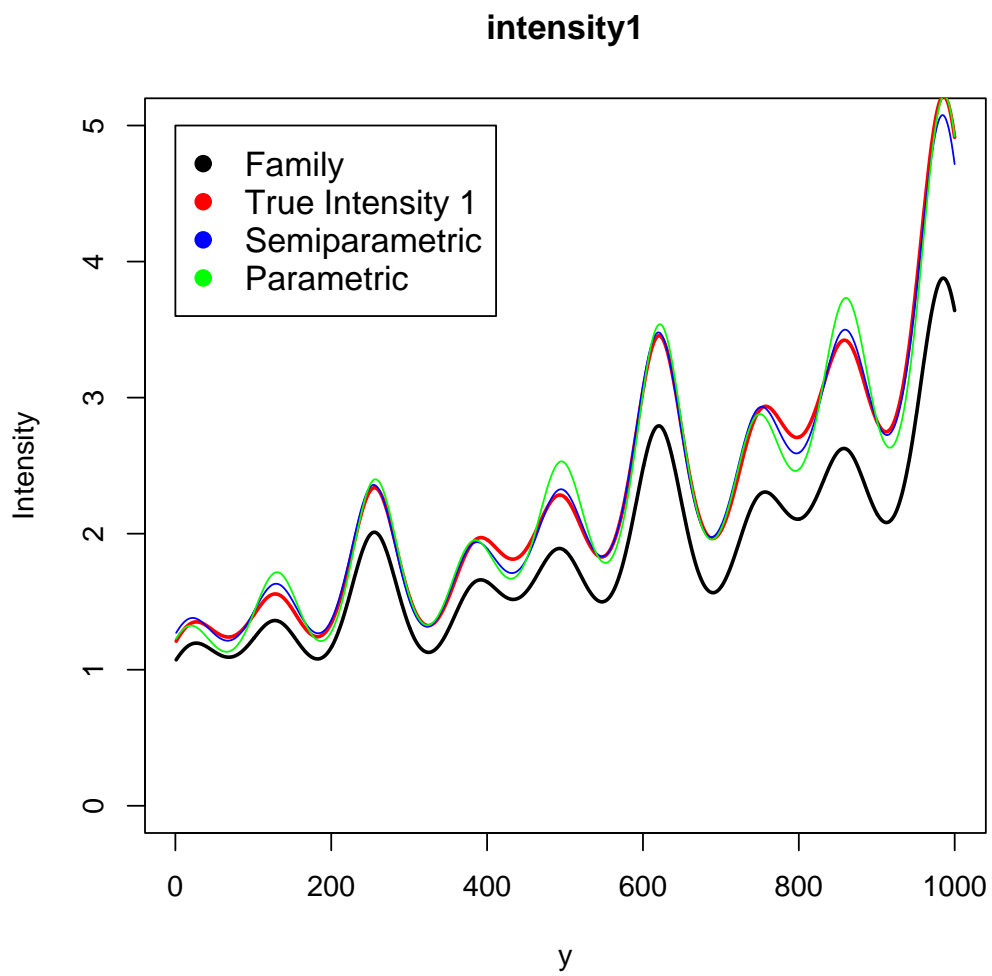


Figure 2.5: intensity 1

Intensity 6

As for Intensity 1 the procedure is carried out for intensity 6 too, with the same hypothesis . The semiparametric estimator performed very very well as you can see in fig. 2.6. Now I want to underline the behaviour and the explanatory capacity of the correction factor plotted in cyan. In the first half time period the true intensity is less tall than the common pattern so the correction factor drive down the parametric estimate. In the second half time period the behaviour is the inverse: Intensity 6 overpasses the common pattern so the correction has to drive up the parametric part and it does it.

When it is needed, the correction factor is active making the estimates closer to true, while preserving the possibility to include a parametric part which is more easy to interpret.

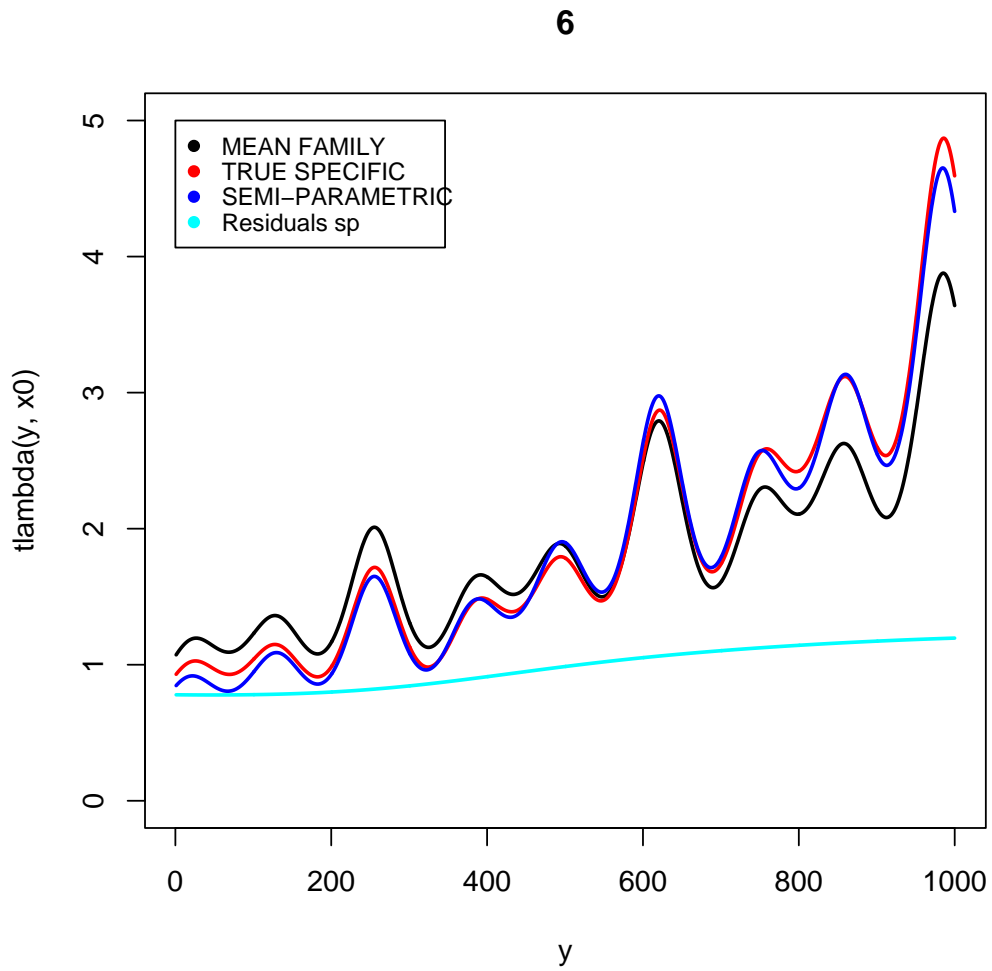


Figure 2.6: intensity 6

Chapter 3

Spatial dependence in the tails of NO_2 distributions: an extreme value theory approach

3.1 Introduction

The chemical compound nitrogen dioxide (NO_2) is a reddish or orange/brown gas with a characteristic sharp, biting odor. It's one of the most prominent air pollutants. Nitrogen dioxide is toxic by inhalation. Symptoms of poisoning (lung edema) tend to appear several hours after one has inhaled a low but potentially fatal dose. Also, low concentrations will anesthetize the nose, thus creating potential for overexposure. Long term exposure to NO_2 at concentrations above $40 - 100 \mu\text{g}/\text{m}^3$ causes adverse health effects. The most important source of NO_2 are internal combustion engines, which emit nitrogen oxides near people. Nitrogen dioxide plays a role in the atmospheric chemistry too, contributing to the formation of acid rains with potentially deep and adverse effects on the ground and aquatic ecosystems.

In Italy, concern about the level of pollution in many cities led to two important policy implementations: the first was the declaration of new legal limits for the tolerance of high concentration levels of a variety of pollutants. The second was the development of an extensive network of monitoring stations with the dual aim of checking the respect

of the legal limits and of providing a database that would permit the study of the behaviour of the pollutants.

The study of the spatial dimension of the distribution of air pollutants could improve the knowledge of the underlying stochastic process. In particular we are interested in modelling the joint spatial risk of the occurrence of extreme cases which are potentially the most dangerous ones. The results of such spatial dependence analysis may improve the spatial prediction or design a better reallocation of monitoring sites.

The basic framework is the estimation and the interpretation of the dependogram (Arbia and Lafratta, 2005 (1)) which is a general measure of nonlinear dependence between two sites restricted to the right tail of their joint bivariate distribution. An improved inference method is proposed here using univariate and bivariate threshold model (Coles and Tawn, 1996 (10)) to make more accurate the estimation of the probability of extreme outcomes which are the targets of our research.

The data we examine have been collected by 7 monitoring stations in Rome over a period of 2 years: 2000 and 2001. The choice of these particular stations is motivated by their spatial localizations: a great portion of Rome is covered and a wide range of distances between sites are represented.

3.2 Spatial Joint Risk

Prediction of the risk of the occurrence of pollutants' extreme cases is one of the biggest problems in the environmental studies. Our approach consists in the use of space and spatial relationship to model such risk avoiding to restrict attention just to linear dependence but also the analysis of multivariate distributions of the process in their whole can be misleading: our interests fall in just a portion of them, that is, the

dangerous cases.

The most popular way of modelling the spatial structure of events is through the notion of spatial correlation and through related instruments as the semivariogram, but the use of such instruments, however, are justified under the implicit assumption that:

- the dependence in all the bivariate marginals of the data generating process is only linear, (hence its study can be restricted to spatial correlation);
- data are distributed as Gaussian (hence linear dependence is the only form of dependence);
- the whole pairwise bivariate marginals of the generating random field are of interest to the analysis.

This is not often the case in environmental studies where phenomena are often non-linear and non-normal. It is certainly not the case when studying extreme environmental events where the interest is restricted to phenomena of dependence in the tails of the distribution.

In studying extreme events it is fundamental the notion of risk: the definition of risk is very general and finds applications in many different fields, we adopt its Cramer's version:

$$Risk_u = \int_u^{\infty} f_Y(y)dy \quad (3.2.1)$$

i.e. the probability of exceeding (or not exceeding) a certain cut-off quantity perceived as dangerous .

This measure is more general than a measure based on the variance of the distribution and taking account of possible asymmetries and other nonnormalities it places a greater emphasis on the tails of the distribution.

In the spatial context, let α be the selected level of risk and let u_i and u_j be the corresponding critical thresholds at sites i and j for a give variable Y : $u_i = F_{Y_i}^{-1}(\alpha)$ and $u_j = F_{Y_j}^{-1}(\alpha)$, we can introduce the concept of *bivariate spatial risk* as the probability of exceeding the dangerous cutoff at site i , and , simultaneously, exceeding the analog threshlod u_j in the other site j :

$$BiRisk(\alpha, i, j) = \int_{u_i}^{\infty} \int_{u_j}^{\infty} f_{Y_i, Y_j}(y_i, y_j) dy_i dy_j \quad (3.2.2)$$

The bivariate spatial risk can be interpreted as proportional to the local risk at site i conditional to $Y_j \geq u_j$:

$$\begin{aligned} BiRisk(\alpha, i, j) &= \int_{u_i}^{\infty} \int_{u_j}^{\infty} f_{Y_i, Y_j}(y_i, y_j) dy_i dy_j \\ &= (1 - F_{Y_j}(u_j)) \int_{u_i}^{\infty} f_{Y_i|Y_j \geq u_j}(y_i) dy_i \\ &\propto \int_{u_i}^{\infty} f_{Y_i|Y_j \geq u_j}(y_i) dy_i \end{aligned} \quad (3.2.3)$$

3.3 Spatial Tail Dependogram

Now using the general concept of *positive quadrant dependence measure* between two random variables which states that in precence of such dependence the following is true:

$$Pr(Y_i \geq y_i \cap Y_j \geq y_j) \geq Pr(Y_i \geq y_i)Pr(Y_j \geq y_j)$$

which in term of BiRisk becomes

$$\int_{u_i}^{\infty} \int_{u_j}^{\infty} f_{Y_i, Y_j}(y_i, y_j) dy_i dy_j - \int_{u_i}^{\infty} f_{Y_i}(y_i) dy_i \int_{u_j}^{\infty} f_{Y_j}(y_j) dy_j$$

and using the 3.2.3:

$$\int_{u_i}^{\infty} f_{Y_i|Y_j \geq u_j}(y_i) dy_i - \int_{u_i}^{\infty} f_{Y_i}(y_i) dy_i$$

Thus, the quantity

$$\xi(\alpha) = \int_{u_i}^{\infty} f_{Y_i|Y_j \geq u_j}(y_i) dy_i - \int_{u_i}^{\infty} f_{Y_i}(y_i) dy_i = F_{Y_i}(u_i) - F_{Y_i|Y_j \geq u_j}(u_i) \quad (3.3.1)$$

is a general measure of nonlinear dependence between site i and site j restricted to the right tail of their joint bivariate distribution. We can index this quantity with the distance $d_{i,j}$ between site i and site j , so we are able to plot $\xi(\alpha, d_{i,j})$ against $d_{i,j}$ and it is what we will refer to as *spatial tail dependogram*.

3.4 Nonparametric Spatial Dependogram Estimation

Arbia and Lafratta (2005,(1)) proposed to estimate $\xi(\alpha, d_{i,j})$ nonparametrically using the empirical distribution function.

$$\hat{\xi}(\alpha, d_{i,j}) = \hat{F}_{Y_i}(\hat{u}_i) - \hat{F}_{Y_i|Y_j \geq \hat{u}_j}(\hat{u}_i) \quad (3.4.1)$$

where \hat{u}_i and \hat{u}_j are the sample quantiles of level α respectively estimated from Y_i and Y_j and

$$\hat{F}_{Y_i}(\hat{u}_i) = \frac{1}{T} \sum_{i=1}^T I_{(-\infty, \hat{u}_i)}(y_{i,t}) \quad (3.4.2)$$

$$\hat{F}_{Y_i|Y_j \geq \hat{u}_j}(\hat{u}_i) = \frac{\frac{1}{T} \sum_{i=1}^T I_{(-\infty, \hat{u}_i)}(y_{i,t}) I_{(\hat{u}_j, \infty)}(y_{j,t})}{1 - \hat{F}_{Y_j}(\hat{u}_j)} \quad (3.4.3)$$

Simulation Study

Arbia and Lafratta (2005,(1)) showed an empirical example based on a simulated dataset. The analysis refers to a dataset sampled from a random field for which the multivariate skew normal distribution (Azzalini and Dalla Valle, 1996, (3)) is used as

the spatial data generating process to allow non-gaussian effects and to create tail dependencies. The simulation is based on four sampled locations distributed in the space producing six distances 3.1 at regular intervals (5, 10, 15, 20, 25, 30).

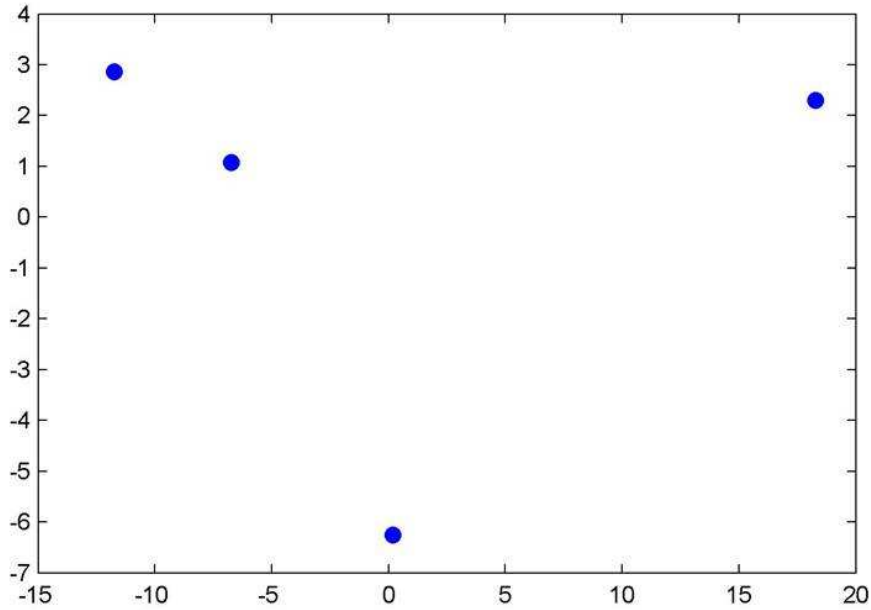


Figure 3.1: Sampled locations

So, the random vector $Y = (Y_1, Y_2, Y_3, Y_4)'$ has four-dimensional skew-normal distribution:

$$f_{Y_1, Y_2, Y_3, Y_4}(y) = 2\phi_4(y; \Omega)\Phi(\theta'y) \quad (3.4.4)$$

where Ω and θ are parameters, $\phi_4(\cdot; \Omega)$ is the four-dimensional Gaussian density having zero mean and covariance matrix Ω and Φ is the univariate Gaussian standard distribution. The drawn random sample is: hourly observations for 5 years: a time series of 43200 observations.

The non-parametric procedure was applied and the estimated spatial tail dependogram

was plotted in fig. 3.2. The plot of the tail correlogram is also provided for a comparison.

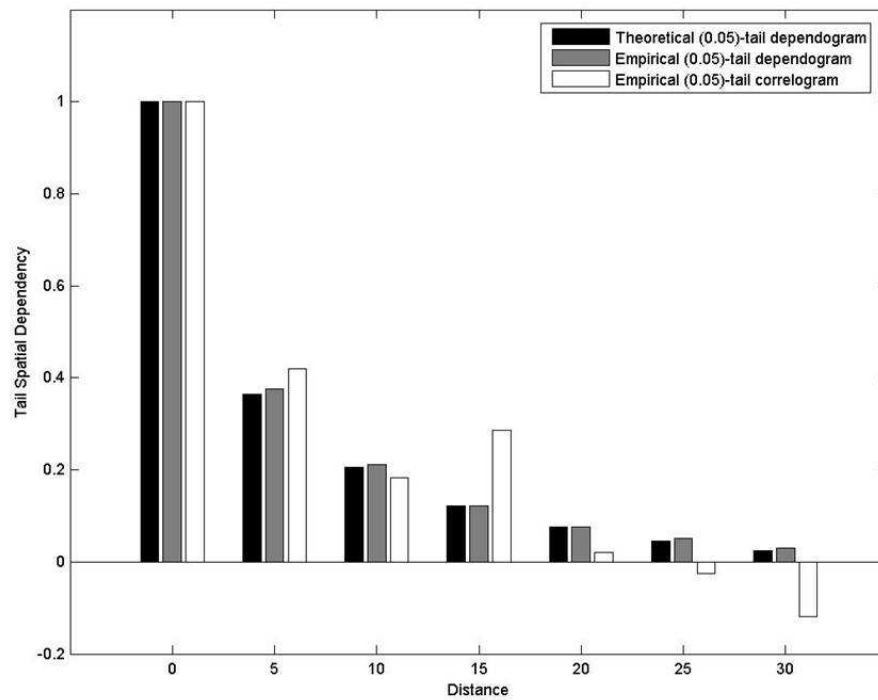


Figure 3.2: Theoretical and empirical dependograms for a probability level of 0.05

Fig. 3.2 shows that the tail correlogram can capture just the linear dependence among extremes, while spatial tail dependogram is a valid measure of spatial dependence restricted to the right tail of the distribution.

We want now to improve di inference framework.

3.5 Parametric Inference for Dependogram: an extreme value theory approach

Handling the second term of the spatial tail dependogram 3.4.1:

$$\begin{aligned}
 F_{Y_i|Y_j \geq u_j}(u_i) &= \frac{\Pr(Y_i \leq u_i \cap Y_j \geq u_j)}{\Pr(Y_j \geq u_j)} \\
 &= \frac{\Pr(Y_i \leq u_i) - \Pr(Y_i \leq u_i \cap Y_j \leq u_j)}{\Pr(Y_j \geq u_j)} \\
 &= \frac{F_{Y_i}(u_i) - F_{Y_i, Y_j}(u_i, u_j)}{1 - F_{Y_j}(u_j)} \tag{3.5.1}
 \end{aligned}$$

which leads to

$$\xi(\alpha, d_{i,j}) = F_{Y_i}(u_i) - \frac{F_{Y_i}(u_i) - F_{Y_i, Y_j}(u_i, u_j)}{1 - F_{Y_j}(u_j)}. \tag{3.5.2}$$

All the probability presents in 3.5.2 are computed in very high value of the random variables: in extreme values. This allow us to impose the inference using the extreme values theory (see Coles S. 2001(8)). So $F_{Y_i}(u_i)$ and $F_{Y_j}(u_j)$ can be modeled as *univariate threshold model*. $F_{Y_i, Y_j}(u_i, u_j)$ as *bivariate threshold model*

3.5.1 Univariate threshold model

To represent the distributional behaviour in the tails, Coles and Twan (9),(10) suggested the use of a point process characterization of multivariate extremes, which can be applied to the spatial case by assuming that each (univariate) component represents, marginally, the polluting process at a given sampled location. Simultaneous estimation of the spatial dependency structure among extremes and of their marginal parameters can thus be obtained by using a maximum likelihood procedure under the assumption of temporal independence.

Some usefull results from Coles, 2001 (8)

Under suitable conditions the random variable

$$M_n = \max \{C_1, \dots, C_n\}$$

converges in distribution, as $n \rightarrow +\infty$, to a member of the Generalized Extreme Value (GEV) family, in the sense that

$$\Pr(M_n \leq c) \approx GEV(c; \eta, \sigma, \xi) \equiv \exp \left\{ - \left[1 + \xi \left(\frac{c - \eta}{\sigma} \right) \right]^{-1/\xi} \right\}, \quad (3.5.3)$$

which is defined on $\{c \in \mathbb{R} : 1 + \xi(c - \eta)/\sigma > 0\}$, with parameters satisfying the constraints $\eta, \xi \in \mathbb{R}$ and $\sigma > 0$

Let us assume that τ_0 represents a threshold high enough to guarantee that, for $\tau > \tau_0$, $\ln(\Pr(C \leq \tau))$ can be approximated in a neighborhood of 1 as follows:

$$\ln(\Pr(C \leq \tau)) \approx \Pr(C \leq \tau) - 1, \quad (3.5.4)$$

Since $\Pr(M_n \leq \tau) = \Pr(C \leq \tau)^n$, applying equations (3.5.3) and (3.5.4) we have

$$1 - \Pr(C \leq \tau) \approx n^{-1} \left[1 + \xi \left(\frac{\tau - \eta}{\sigma} \right) \right]^{-1/\xi}$$

and thus

$$\begin{aligned} \Pr(C > \tau | C > \tau_0) &\approx \frac{1 - \Pr(C \leq (\tau - \tau_0) + \tau_0)}{1 - \Pr(C \leq \tau_0)} \\ &= \left[1 + \xi \left(\frac{\tau - \tau_0}{\tilde{\sigma}} \right) \right]^{-1/\xi}, \end{aligned} \quad (3.5.5)$$

with $\tilde{\sigma} = \sigma + \xi(\tau_0 - \eta)$. This is equivalent to state that, conditional on $C > \tau_0$, the distribution of $C - \tau_0$ belongs to the generalized Pareto family, with parameters ξ and $\tilde{\sigma}$

Formula (3.5.5) suggests that, for every $\tau > \tau_0$, we can state that:

$$\Pr(C > \tau) \simeq \zeta_{\tau_0} \left[1 + \xi \left(\frac{\tau - \tau_0}{\tilde{\sigma}} \right) \right]^{-1/\xi}, \quad (3.5.6)$$

where $\zeta_{\tau_0} = \Pr(C > \tau_0)$, ξ and $\tilde{\sigma}$ need to be estimated. Applying the Maximum Likelihood principle leads to the sample proportion of points exceeding τ_0 as an estimator of ζ_{τ_0} , and to estimates of ξ and $\tilde{\sigma}$ which can be obtained numerically as discussed by Davison and Smith (1990, (14)).

advanced inference for dependogram 1

Just applying previous results to estimate dependogram, we obtain:

$$1 - F_{Y_i}(u_i) = \Pr(Y_i > u_i) \approx \zeta_{u_0} \left[1 - \xi \left(\frac{u_i - u_0}{\tilde{\sigma}} \right) \right] \quad (3.5.7)$$

and the maximum likelihood based inference for the parameters is as in Coles (2001, (8)).

3.5.2 Bivariate threshold model

Modelling the bivariate joint distribution in terms of threshold model is instead not so immediate.

Let us assume for a moment that both the components of the random vector (C_j, C_l) have marginal unit Fréchet distributions, i.e. $\Pr(C_j \leq c) = \Pr(C_l \leq c) = \exp(-c^{-1}) \equiv \psi(c)$, $c > 0$. Under such condition, if the vector of componentwise maxima rescaled by n^{-1} ,

$$\mathbf{M}_n^* = \begin{pmatrix} n^{-1}M_{j,n} & n^{-1}M_{l,n} \end{pmatrix},$$

converges in distribution to a non-degenerate function, say $G(c_j, c_l)$, then G can be characterized as follows:

$$G(c_j, c_l) = \exp(-V(c_j, c_l)), \quad (3.5.8)$$

where the function V is such that

$$V(c_j, c_l) = 2 \int_0^1 \max\left(\frac{w}{c_j}, \frac{1-w}{c_l}\right) dH(w), \quad (3.5.9)$$

H being a distribution on $[0, 1]$ having mean $1/2$. Such result thoroughly specifies the class of bivariate limit distributions for

$$\mathbf{M}_n = \begin{pmatrix} M_{j,n} & M_{l,n} \end{pmatrix}.$$

In fact, equation (3.5.9) constrains V to be homogeneous of order -1 , which means that, for every $\delta > 0$,

$$V(\delta^{-1}c_j, \delta^{-1}c_l) = \delta V(c_j, c_l),$$

so that

$$F_{C_j, C_l}(c_j, c_l)^n = \Pr(M_{j,n} \leq c_j \cap M_{l,n} \leq c_l) \approx G(n^{-1}c_j, n^{-1}c_l) = G(c_j, c_l)^n,$$

and hence

$$F_{C_j, C_l}(c_j, c_l) \approx G(c_j, c_l). \quad (3.5.10)$$

The approximation in (3.5.10) holds true if both C_j and C_l are unit Fréchet distributed. Nevertheless, an analog result can be stated for arbitrary marginal distributions of (C_j, C_l) . More precisely, it is always possible to apply, for every $j \in J$, the transformation

$$y = \theta_j(c) = \psi^{-1}(F_{C_j}(c)), \quad (3.5.11)$$

so that the vector

$$(Y_j, Y_l) = (\theta_j(C_j), \theta_l(C_l)),$$

has, by construction, marginal unit Fréchet distributions. Hence, we would have, by (3.5.10),

$$F_{Y_j, Y_l}(y_j, y_l) \approx G(y_j, y_l),$$

and thus

$$\begin{aligned} F_{C_j, C_l}(c_j, c_l) &= \Pr(\theta_j(C_j) \leq \theta_j(c_j) \cap \theta_l(C_l) \leq \theta_l(c_l)) \\ &\approx G(\theta_j(c_j), \theta_l(c_l)). \end{aligned} \quad (3.5.12)$$

We can now observe that, for a suitable threshold $\tau > \tau_0$, the marginal distribution of Y , can be approximated using equation (3.5.6), after marginal estimation of the corresponding parameter set $(\zeta_{\tau_0}, \tilde{\sigma}, \xi)$:

$$F_Y(\tau) \approx 1 - \zeta_{\tau_0} \left[1 + \xi \left(\frac{\tau - \tau_0}{\tilde{\sigma}} \right) \right]^{-1/\xi}.$$

If such approximations are plugged in equation (3.5.11), we obtain,

$$\theta(\tau) = - \left(\ln \left(1 - \zeta_{\tau_0} \left[1 + \xi \left(\frac{\tau - \tau_0}{\sigma} \right) \right]^{-1/\xi} \right) \right)^{-1}, \quad (3.5.13)$$

so that the joint distribution $F_{Y_j, Y_l}(y_j, y_l)$ can be approximated on the region $\{(y_j, y_l) : y_j > \tau, y_l > \tau\}$ using equations (3.5.11) and (3.5.12).

advanced inference for dependogram 2

The last results allow us to modeling the bivariate threshold model to estimate the bivariate probability present in the dependogram.

If we apply the following transformation

$$Z_i = - \left(\ln \left(1 - \zeta_{u_0} \left[1 + \xi \left(\frac{u_i - u_0}{\sigma} \right) \right]^{-1/\xi} \right) \right)^{-1} \quad (3.5.14)$$

The Z 's become Frechet's marginals, and the following result holds:

$$F_{Y_i, Y_j}(u_i, u_j) = \exp(-V(Z_i, Z_j)), \quad (3.5.15)$$

where V is a homogeneous function of order -1 . In this work I used the *logist model* to characterize V as:

$$V(z_i, z_j) = \left(z_i^{-1/\rho} + z_j^{-1/\rho} \right)^\rho \quad (3.5.16)$$

where $0 \leq \rho \leq 1$ measures the strength of the dependence and is an extra parameter to be estimated.

As for univariate threshold model, here a maximum likelihood approach is carried out (see Coles, 2001 (8)) to estimate the parameter ρ .

3.5.3 Choice of the threshold

What is the threshold that let be true the GPD approximation for extreme value? How to choose it?

An explanatory tool is available to help us: the *Mean Residual Life Plot*. It consists in the plot of the pairs:

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \right) : u < x_{max} \right\} \quad (3.5.17)$$

where $x_{(i)}$ is the i -th observation exceeding the threshold u and n_u is the number of

observations exceeding the threshold u . The mean residual life plot should be approximately linear in u if we trespass the threshold where the Generalized Pareto distribution provides a valid approximation to the excess distribution.

3.6 Empirical Application: the NO₂ case in Rome

In this section we will illustrate the methodology described in the previous section by making use of a real data set referred to the distribution of NO₂ in Rome. The obtained estimates of spatial tail dependogram is plotted at varies distances to detect spatial regularities in the distribution of extremes of NO₂.

NO_2 hourly data have been collected by seven monitoring stations in Rome in the years 2000 and 2001. We consider just the winter months for homogeneity for a total of 4906 observations. The spatial locations of the seven monitoring sites in Rome are plotted in fig. 3.3 generating 21 distances.

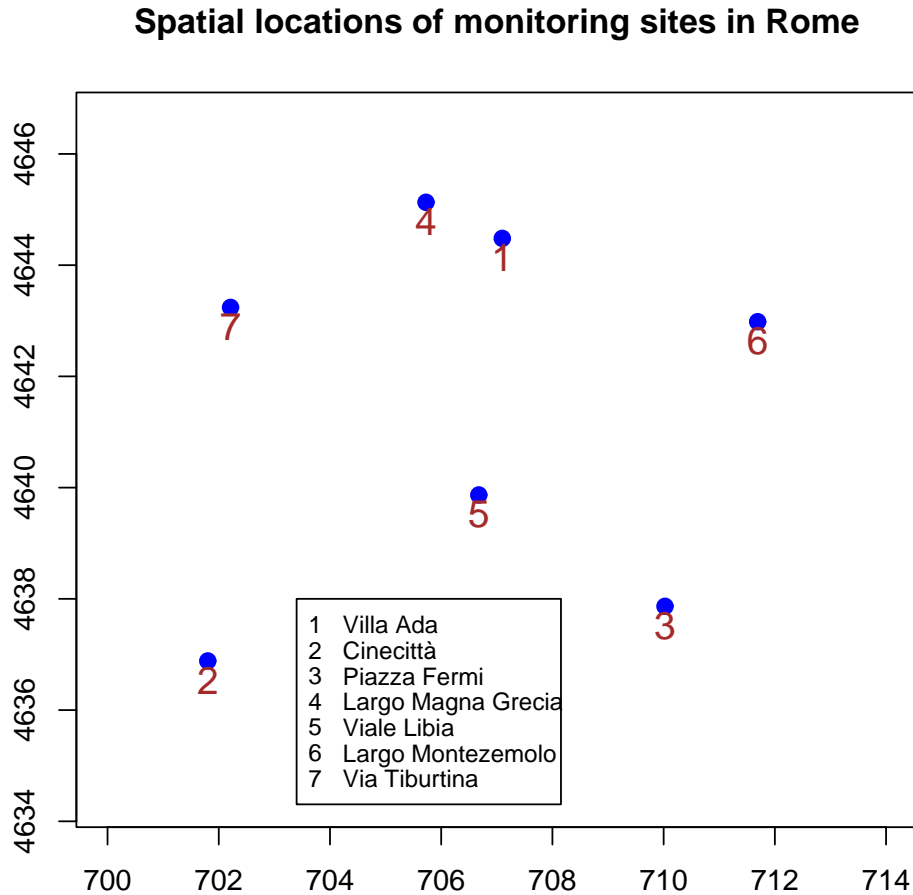


Figure 3.3: Spatial locations of the seven monitoring sites in Rome

So we are considering seven random variable Y_1, \dots, Y_7 whose extremes will be modelled with univariate and bivariate threshold models. The first step consists in the choice of the threshold for each of the seven variables. We need to plot the mean

residual life plot for each monitoring site and check for the value after which the plot becomes linear in u .

The mean residual life plot for site 3 - *Piazza Fermi* - is showed in fig. 3.4 and suggests the use of the value 156 as a critical threshold which correspond to the 97-th percentile. The critical thresholds for the others sites are smaller than the respective 97-th percentiles, so, we can assume, without risk, a common threshold in terms of percentiles, being sure that it is the biggest one for all the seven monitoring sites. So $u_{0,i} = F_i^{-1}(0.97)$ for $i = 1, \dots, 7$.

Now we are able to make inference for each of the seven random variables, modeling their extremes with univariate threshold model as in 3.5.7. The estimated parameters are reported in table 3.1.

site	1	2	3	4	5	6	7
threshold	92,64	122,58	156,71	136,31	139,08	138,08	147,78
$\hat{\sigma}$	10,830	22,989	21,067	16,476	15,699	12,123	14,064
$s.e.(\hat{\sigma})$	1,516	2,775	2,470	1,966	1,861	1,363	1,621
$\hat{\xi}$	0,327	-0,193	-0,063	0,036	0,093	-0,002	0,103
$s.e.(\hat{\xi})$	0,116	0,090	0,084	0,087	0,086	0,077	0,081

Table 3.1: Univariate Threshold Model Parameter Estimation

Then for each pair of monitoring sites I estimated the bivariate threshold model; after done the transformation 3.5.14 and implemented the logistic model, only the parameter ρ has to be estimated. And result are reported in tabel 3.2

At this stage we have to estimate the spatial tail dependogram. Let us fix the dangerous probability at the level, say $\alpha = 0.02$. We will obtain for each monitoring site the dangerous cutoff at which compute the univariate and bivariate probabilities present in the last formula of the dependogram 3.5.2. The probabilities are reported in Table 3.3 and in table 3.4. All we need now is to combine these univariate e bivariate probabilities as showed in the formula 3.5.2 for each pairs of monitoring sites which

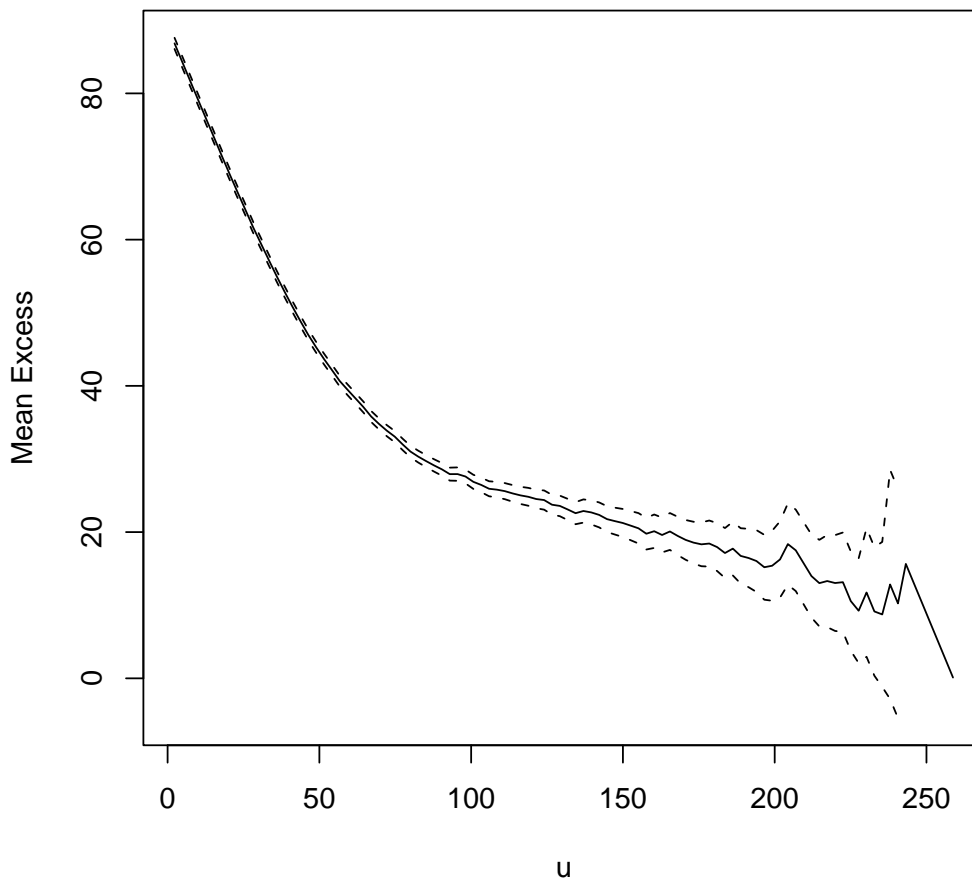


Figure 3.4: Mean Residual Life Plot: Site 3 Piazza Fermi

are at distance $d_{i,j}$. The spatial tail dependogram is computed and linked to the distances as showed in Table 3.5, and finally plotted in Fig.3.5 and compared with the spatial correlogram. The analysis of the dependence restricted to the right tail of the distributions through the spatial tail dependogram thus shows a certain decreasing trend with distance. Although this is only a modest trend, the spatial correlogram (that considers all values in the bivariate distribution and not only those in the upper tail) displays the opposite trend which is contrast with any geographical intuition. Values

site i	site j	$\hat{\rho}$	$s.e.(\hat{\rho})$
1	2	0,836	0,022
1	3	0,824	0,023
1	4	0,726	0,025
1	5	0,754	0,026
1	6	0,816	0,023
1	7	0,852	0,022
2	3	0,789	0,023
2	4	0,748	0,024
2	5	0,840	0,022
2	6	0,872	0,020
2	7	0,805	0,023
3	4	0,715	0,025
3	5	0,868	0,021
3	6	0,804	0,023
3	7	0,866	0,020
4	5	0,759	0,025
4	6	0,788	0,024
4	7	0,811	0,023
5	6	0,837	0,023
5	7	0,780	0,025
6	7	0,780	0,023

Table 3.2: Bivariate Threshold Model Parameter Estimation

are more similar if they are distant than if they are close together. Thus the spatial tail dependogram seems to provide a more sensible description of reality. Anyway the analysis was affected by the presence of just one monitoring site at very short distance and just one monitoring site at very large distance. A better reallocations of them is needed.

Site	Threshold	$P(y_i > u_i)$ (Univariate threshold models)	$F_i(u_i)$
Site 1	97,46	0,019814	0,980186
Site 2	130,87	0,02065	0,97935
Site 3	164,40	0,020742	0,979258
Site 4	143,21	0,019803	0,980197
Site 5	146,43	0,018967	0,981033
Site 6	142,92	0,020131	0,979869
Site 7	154,02	0,019444	0,980556

Table 3.3: Univariate estimated probabilities at dangerous level equal to 0.02

site i	site j	$P(y_i > u_i, y_j > u_j)$ (Bivariate threshold models)	$F_{i,j}(u_i, u_j)$
1	2	0,00438	0,96392
1	3	0,00467	0,96412
1	4	0,00690	0,96728
1	5	0,00610	0,96732
1	6	0,00482	0,96487
1	7	0,00385	0,96459
2	3	0,00566	0,96427
2	4	0,00652	0,96607
2	5	0,00417	0,96456
2	6	0,00348	0,96270
2	7	0,00509	0,96500
3	4	0,00729	0,96675
3	5	0,00348	0,96377
3	6	0,00522	0,96435
3	7	0,00359	0,96340
4	5	0,00599	0,96722
4	6	0,00549	0,96556
4	7	0,00485	0,96561
5	6	0,00419	0,96509
5	7	0,00546	0,96705
6	7	0,00563	0,96605

Table 3.4: Bivariate estimated probabilities at dangerous level equal to 0.02

site i	site j	$\xi(\alpha, d_{i,j})$	$d_{i,j}(km)$
1	4	0,3287	1,5176
3	5	0,1628	3,9012
4	7	0,2298	3,9907
1	5	0,3019	4,6326
1	6	0,2195	4,8312
1	7	0,1781	5,0415
4	5	0,2960	5,3475
3	6	0,2387	5,3821
5	7	0,2617	5,5949
2	5	0,1993	5,7148
5	6	0,1891	5,9037
4	6	0,2531	6,3395
2	7	0,2411	6,3702
1	3	0,2055	7,2339
2	3	0,2521	8,2799
3	4	0,3476	8,4409
2	4	0,3087	9,1330
1	2	0,1923	9,2613
3	7	0,1638	9,4827
6	7	0,2693	9,4829
2	6	0,1523	11,6184

Table 3.5: Estimated spatial tail dependogram at dangerous level equal to 0.02

3.7 Conclusions

In this chapter we considered the problem of estimating the spatial dependence between neighbouring location by looking at what happens in the left tails of the bivariate distribution. In this way we restrict ourselves to those cases where a certain dangerous threshold is trespassed simultaneously in two closeby locations.

The analysis reported here considered the exploratory tool termed *spatial tail dependogram* introduced by Arbia and Lafratta (2005, (1)), but, rather than considering the non-parametric estimator suggested by the two authors, here we considered a fully parametric likelihood-based version of it.

The empirical example reported in the paper was aiming at showing how the new estimation procedure works in practice. It focused on the spatial distribution of NO₂ in Rome as reported in seven monitoring stations in the years 2000 and 2001. The spatial tail dependogram was estimated parametrically on the basis of the empirical data and the results were compared with those obtained by using the more traditional spatial correlation.

The main finding is that, while the spatial correlogram shows a counter-intuitive increasing behaviour with respect to distance, the spatial tail dependogram conversely displays a more intuitive decreasing behaviour which accords with the first *law of geography* (Tobler, 1970, (25)).

Obviously more work is needed in this field to overcome some of the limitations of the present analysis. An important aspect concerns the empirical analysis that should be run on a dataset larger with respect to the spatial dimension. In our case we had only 7 observations giving rise to 21 possible distances. So both the correlogram and the spatial tail dependogram were interpolated to only 21 points. A larger dataset would allow a better interpolation and hence more grounded conclusions when comparing the

two plots. However it is still difficult to avail a large environmental database related to a large number of monitoring sites and which is also simultaneously rich in terms of the temporal replication that are required to activate our estimation procedure.

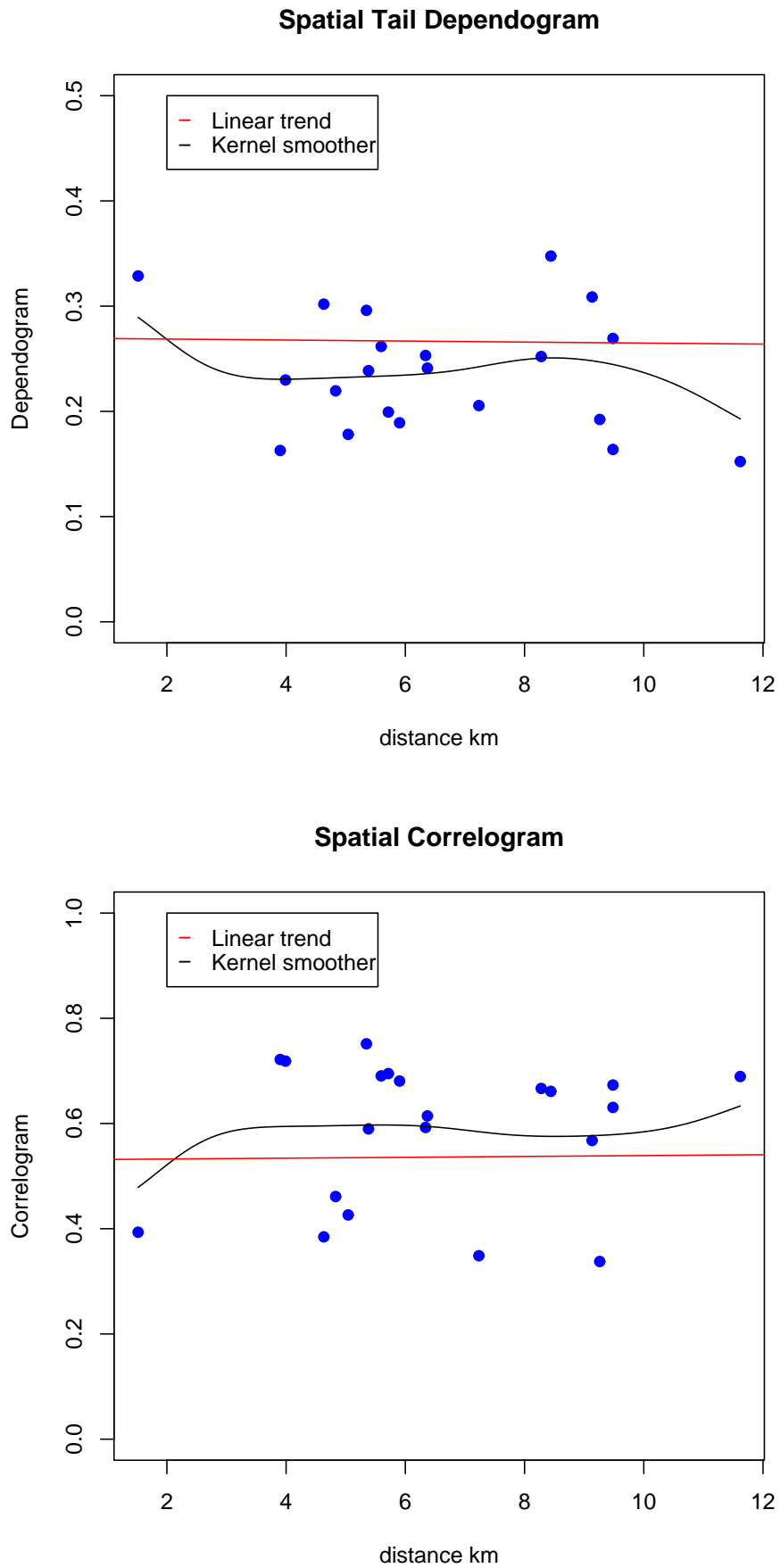


Figure 3.5: Spatial Tail Dependogram and Spatial Correlogram

Bibliography

- [1] Arbia G., Lafratta G. 2005 Exploring Nonlinear Spatial Dependence in the Tails. *Geographical Analysis*, **37**: 423-427
- [2] Arbia G., Lafratta G., Simeoni C. 2006 Spatial Sampling Plans to Monitor the 3-D Spatial Distribution of Extremes in Soil Pollution Surveys. *Computational Statistics and Data Analysis*, **forthcoming**
- [3] Azzalini A., Dalla Valle A. 1996 The multivariate skew-normal distribution. *Biometrika*, **83(4)**: 715-726
- [4] Bartlett M. S. 1964. The spectral analysis of twodimensional point processes. *Biometrika*, **51**: 299-311.
- [5] Berman M., Diggle P. J. 1989. Estimating Weighted Integrals of The Second-order Intensity of a Spatial Point Processes. *Journal of Royal Statistical Association Ser.B*, **51**: 81-92.
- [6] Bowman A. W., Azzalini A. 1997. Applied Smoothing Techniques for Data Analysis. *Oxford University Press, New York*.
- [7] Breschi S., Lissoni F. 2004. Knowledge Networks from Patent Data: Methodological Issues and Research Targets. *CESPRI WP N° 150*.

-
- [8] Coles S. G., 2001 An introduction to statistical modeling of extreme values. *Springer*
- [9] Coles S. G., Tawn J. A. 1991 Modelling extreme multivariate events. *Journal of the Royal Statistical Society, B Series*, **53(2)**: 377-392
- [10] Coles S. G., Tawn J. A. 1996 Modelling extremes of the rainfall process. *Journal of the Royal Statistical Society, B Series*, **58(2)**: 329-347
- [11] Cox D. R., Lewis P. A. W. 1966. The statistical analysis of series of events. *Methuen, London*.
- [12] Cox D. R. 1955. Some statistical methods related with series of events. *Journal of the Royal Statistical Society, B Series*, **17**: 129-164
- [13] Daley D. J., Vere-Jones D. 1972. A summary of the theory of point processes. *Stochastic Point Processes*, Wiley, New York, 299-383.
- [14] Davison A. C., Smith R. L. 1990 Models for exceedances over high thresholds. *Journal of the Royal Statistical Society, B Series*, **52(3)**: 393-442
- [15] Diggle P. J. 1985. A kernel method for smoothing point process data *JRSS C Applied Statistics*, **34**: 138-147.
- [16] Diggle P. J. 2003. Statistical analysis of spatial point patterns. *Oxford University Press, New York*.
- [17] Diggle P. J., Zheng P., Durr P. 2005. Nonparametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK. *JRSS C Applied Statistics*, **54**: 645-658.

-
- [18] Hjort N. L., Glad I. K. 1995. Nonparametric Density Estimation with a Parametric Start *The Annals of Statistics*, **23**: 882-904
- [19] Jaffe A. B., Trajtenberg M., Fogarty M.S. 2000. Knowledge spillovers and patent citations: evidence from a survey of inventors. *the American economic Review*.
- [20] Jaffe A. B., Trajtenberg M., Henderson R. 1993. Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations. *The Quarterly Journal of Economics*, **108**: 929-952.
- [21] Kelsall J. E., Diggle P. J. 1998. Spatial variation in risk of disease: a nonparametric binary regression approach. *JRSS C Applied Statistics*, **47**, 559-573.
- [22] Ripley B. D. 1977. Modelling Spatial Patterns. *Journal of Royal Statistical Association Ser.B*, **39**, 172-212.
- [23] Silverman B. W. 1986. Density estimation for statistics and data analysis. *Chapman&Hall*.
- [24] Thompson P., Fox-Kean M. 2002 . Patent Citations and the Geography of Knowledge Spillovers: a Reassessment. *the American economic Review*.
- [25] Tobler W. 1970 . A computer model simulation of urban growth in the Detroit region. *Economic Geography*,**42(2)**, 234-240.
- [26] Wand M. P., Jones M. C. 1995. Kernel Smoothing. *Chapman&Hall*.