Statistical modelling of agricultural data in a world of big data

Giuseppe Arbia, Università Cattolica del Sacro Cuore, Roma

Introductory remarks

"From the dawn of civilization until 2003 humankind generated 5 exabytes of data while now we produce 5 exabytes every two days, with an accelerating pace doubling every 40 months" (Google's CEO, Schmidt, 2010).

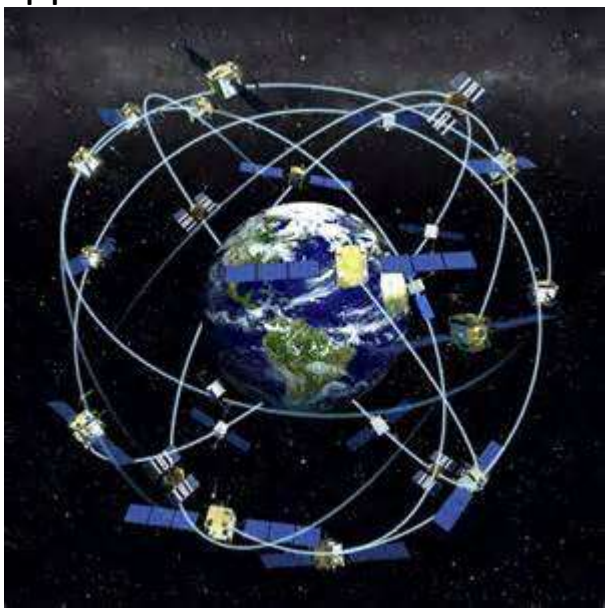
A formidable explosion of data collection and diffusion in all areas of human society !

Researchers are becoming more and more aware of the *"big data"* phenomenon in many scientific fields and of the need to properly manage it combining the tools offered by statistics, probability, mathematics, computer engineering and informatics.

Definition of big data analysis problem: It emerges when the standard procedures fail due to the data dimension (example later)

Introductory remarks

Due to the increased human ability to acquire detailed information through sophisticated technical devices (like Global Positioning Systems, high-resolution remote sensing, other positioning devices), and store them in dedicated Geographical Information Systems (GIS), most of the data automatically generated and collected continuously by public and private institutions are geo-referenced. Upcoming satellites and the introduction innovative positional devices will further improve some applications.

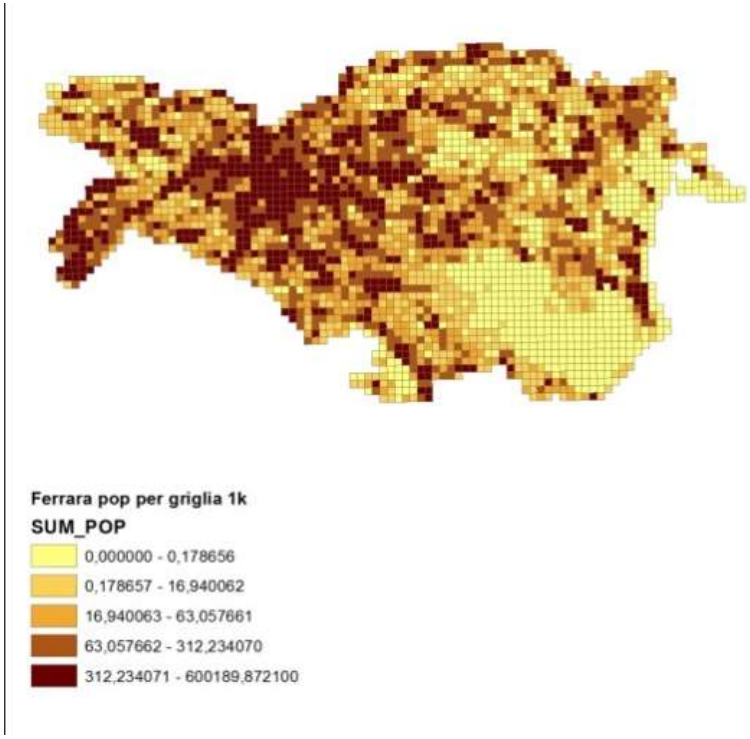


Introductory remarks

Example 1: Agriculture This revolution involves all fields of scientific knowledge, but perhaps it is fair to say that the most relevant applications are in the agricultural and agro-environmental fields (including environmental security, meteorology, geology, oceanography, climate changes) where monitoring and forecasting is critical for policy makers

Introductory remarks

Example 2: Urban/rural structure Eurostat encourages the use of regular grid to analyze and compare population distributions. Even if the 1000 mt grid is the one officially used by Eurostat, the Italian National Statistical System has produced with ISPRA a basic regular grid population distribution of 20mt-by-20 mt squares.

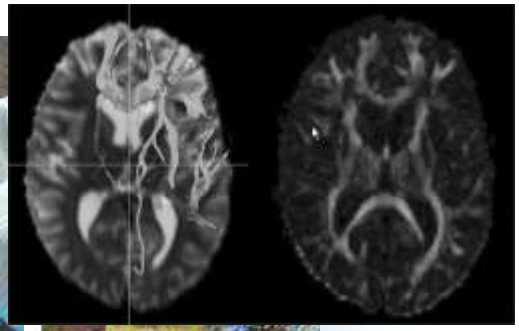


*Population distributions in the 2,633 squares of the 1km-by-*1km grid of Ferrara.

Introductory remarks

Other examples:
Medical computer images, genome mapping, economics, social network analysis, epidemiology, health planning, criminology, archaeology, microbiology, security, migration, transportation, urban traffic, retail transaction, sport, e-commerce, geomarketing, astronomy, military applications, web-generated data ...





Introductory remarks

The computational issues associated to statistical analysis of spatial data were relevant in the 70's.

But even nowadays the burden of calculation can become unbearably demanding in terms of:

- computing time
- computer storage
- accuracy

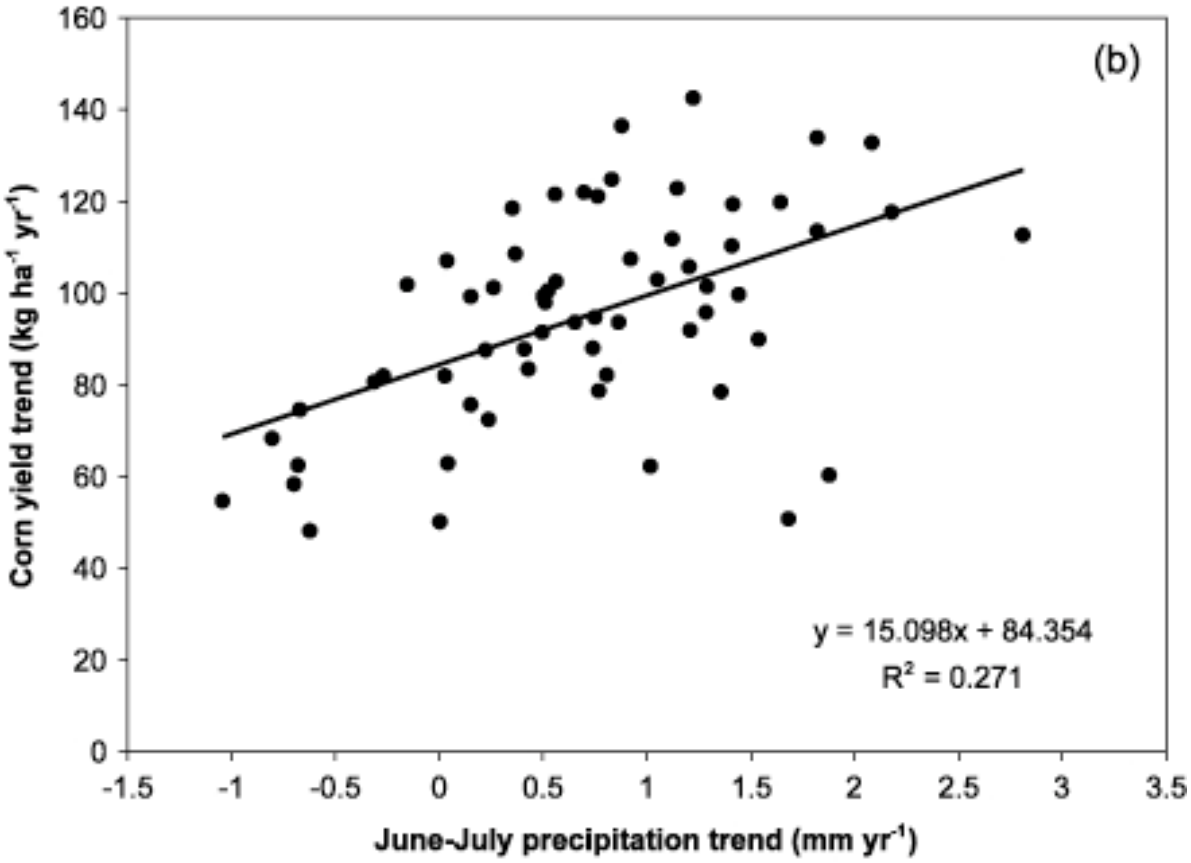
because the availability of very large databases is also increasing.

In the future the gap between data availability and computer ability to treat them will further increase rather than diminish.

A BIG DATA Scientific challenge: *in the absence of appropriate statistical and computer programming techniques, we will most probably find ourselves in the future with massive quantities of spatial data that we will be simply unable to treat timely and accurately.*

The case of spatial linear regression

With spatial data the traditional way of estimating a relationship between variables (a linear regression) lead to **wrong inferential conclusions.**



The case of spatial linear regression

Consequences can be dramatic in many fields:

Example: Measuring the impact of agricultural programs in plant disease control

Neglecting the fact that a policy spills its effects in the neighborhood, we may end up considering a policy to have a significant impact when, in fact, it is ineffective, or, conversely we may end up rejecting policies that were potentially effective.

In many instances, the speed with which we treat massive quantities of spatially dependent data in an accurate way is of vital importance
e. g.

- in environmental disasters management
- in plant disease diffusion control and prevention

Spatial linear regression

Approximations

Ord's (1975) **eigenvalue approach**. Inaccuracte already for sample size of the order of n = 2000

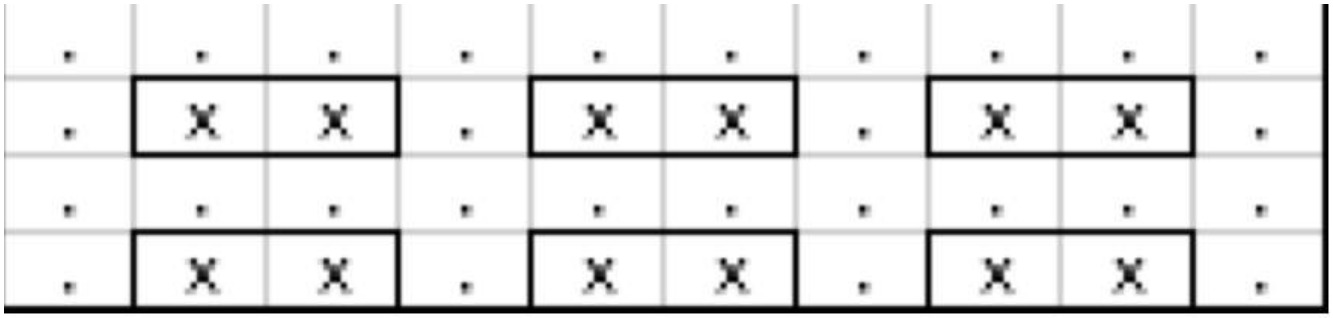
Other approximations: Mead (1967), Whittle(1954), Smirnov and Anselin (2001), Griffith (2000, 2004), Pace and LeSage (2004), Pace and Zou (2000), Pace (1997), Pace and Barry (1997).

Other modelling solutions:

1. GM approach Kelejian and Prucha (1998)
2. Matrix exponential spatial specification *(MESS)* Lesage and Pace (2007)
3. Unilateral approximation (Arbia et al, 2013)
4. Composite likelihood (Arbia, 2014)

A composite likelihood solution (Arbia, 2014)

If we have a large spatial sample (e. g. an agricultural satellite image)



We sample pairs of neighboring observations to retain the spatial information

With this method

- The computing time is ZERO !! whatever is the data dimension
- The storage required is also ZERO
- The accuracy has been proved to be very good in experimental cases

Further possible alternatives
1) Recursive strategies

Instead of estimating a model considering all information simultaneously, we can build it as a mechanism that progressively updates the status using only a smaller amount of information

Examples of this modelling strategy:

- Kalman filters (Kalman, 1960)
- state-space representation (Durbin *and* Koopman, 2001),

The common recursive nature of these algorithms is such that they only use the present input measurements together with the previously calculated state of a system, while no additional past information is necessary.

This leads to a dramatic reduction of the calculations needed when analyzing very large datasets.

A similar approach can be exploited with geo-located dynamic data: by restricting the interest to only local input measurements, together with information on previously calculated local states of the phenomenon, we are able to update the state of the system without needing to consider all spatial observations simultaneously.

Further possible alternatives
2) Multilevel Decomposition

ollowing a multilevel approach we can build up different models at different levels of geographical aggregation, carrying out part of the analysis at the coarser level (with a smaller sample size) thus restricting the computation requested for the finer level based on a larger sample size. Multilevel models are well known (Goldstein, 2003), but, so far, they did not consider the specificity of spatio-temporal data.

Future challenges

My thesis today: *In the absence of appropriate statistical and computer programming techniques, we will probably find ourselves in the future with massive quantities of spatial data that we will be simply unable to treat accurately*.

However, even if in the future the development of computer capabilities could overcome the current limitations of statistical methods, the big data revolution represents a historic occasion offered to statisticians to totally re-think all spatial methodology from its very root.

In the last 30 years we have observed an evolutionary development of methods for spatial data, however, the specificity created by the increasing availability of large geo-coded information urges a revolutionary discontinuity favoring the **outbreak of entirely new statistical models**, methods, techniques and software architectures to answer the new research questions and the emerging challenges.

This will require a progressive shift of the emphasis :

- From sampling errors to non-sampling errors
- From discrete space to continuous space

From sampling to non-sampling errors

In the past we observed a progressive shift from census to sample surveys.

In the future we will observe the opposite.

Data will be more and more collected through indirect sources and the role of official statistics will be less and less collecting data, and increasingly:

- data validation
- data merging
- data quality
- missing data
- missing items
- measurement errors
- locational errors
- data masking for confidentiality

Concluding remarks

We are at the eve of a revolutionary step in the field of spatial data analysis with important bearings on agricultural and rural planning.
This step requires skills in:

- statistics,
- probability
- geography,
- data science,
- computer engineering,
- informatics

In order to develop:

- new statistical theories,
- spatial data analysis methodologies,
- computational techniques,
- visual analytical,
- applications

to treat massive and complex spatial and spatio-temporal datasets in an **accurate** and **efficient** way, acquire new insights and better support individual decisions and policy making.

Thank you for your attention giuseppe.arbia@rm.unicatt.it