ENV-444

Exploratory data analysis in environmental health

Stéphane Joost & Mayssam Nehme

Introduction

Moodle: https://go.epfl.ch/edenv

ENV-444

Exploratory spatial data analysis in environmental health

Stéphane Joost & Mayssam Nehme

Introduction

Moodle: https://go.epfl.ch/edenv

In summary

- This course presents how to apply exploratory methods to health data with a geospatial component
- Application to georeferenced health data, comparison with environmental information
- Theoretical part: ex-cathedra lecture(s)
- Practical part:
  - Individual exercises on your computers
  - Collective work to elaborate a semester project (scientific article writing)

Teachers

- Dr Stéphane Joost, Dr Mayssam Nehme
- Assistant: Noé Fellay
- Teaching assistant: Vacat



- Dr Nehme is the head of the Unit of Population Epidemiology (UEP), Service of Primary Care Medicine (SMPR), Geneva University Hospitals (HUG)

Outline for today

1. Different components of the course and their articulation
2. Situate exploratory spatial data analysis between spatial analysis, data analysis, environmental engineering and statistics
3. John Tukey: exploratory and confirmatory analysis in statistics
4. Organization of the course, requirements

This document is a reference as regards requirements!

Different components of the course

1. Exploratory (Spatial) Data Analysis (EDA - ESDA)
   Appropriate statistics
2. Cognitive aspects for geodata exploration
3. Population epidemiology
   Medical cohorts
4. Effects of environmental conditions on health
   Approach not necessarily geospatial (metabolic syndrome)
5. Spatial epidemiology
   Exploratory Spatial Data Analysis (ESDA) in Environmental health
6. Scientific paper writing

Statistics

Cognition

Population health

Environment x health

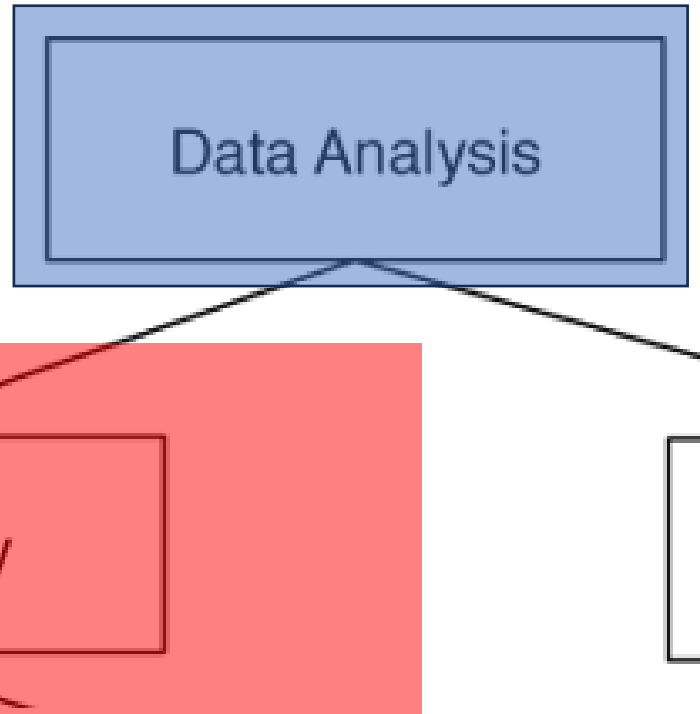
Investigate health & place relation-ship

Type and content of teaching

- ENV-444 is not a technical course in statistics
- It is about a particular sub-domain of statistics: exploratory data analysis and it use in a geospatial context
- It is more on the GIS/mapping/visualization than on the statistical side
- It is about how to combine interactively and dynamically geospatial and statistical tools
- It is also about how to use human cognitive capacities to detect signals

Data analysis

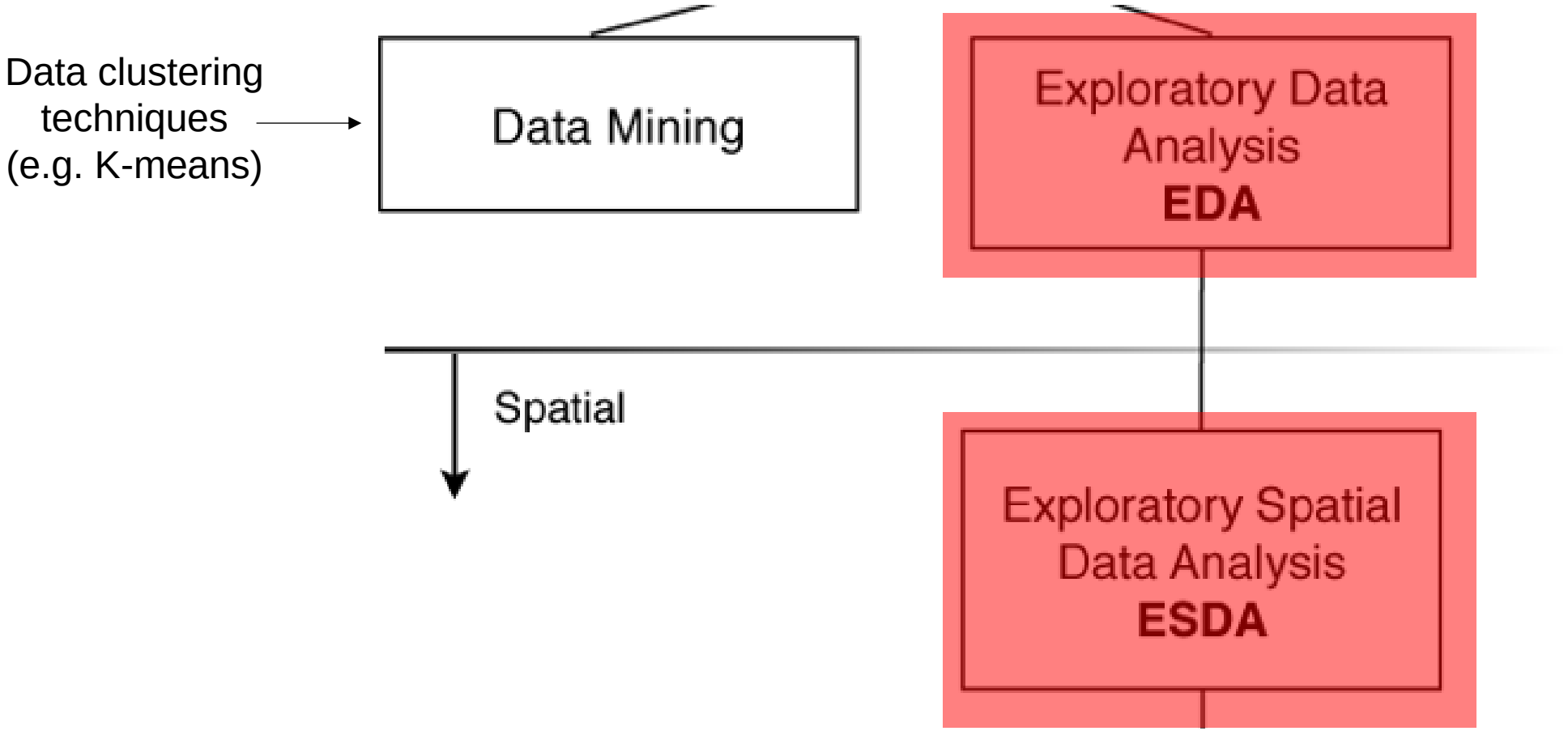
"Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data." (Tukey, 1961)



- No single a priori research question
- The data alone formulate questions…
- …by means of descriptive statistics, exploratory tools
- Favors the formulation of many working hypotheses

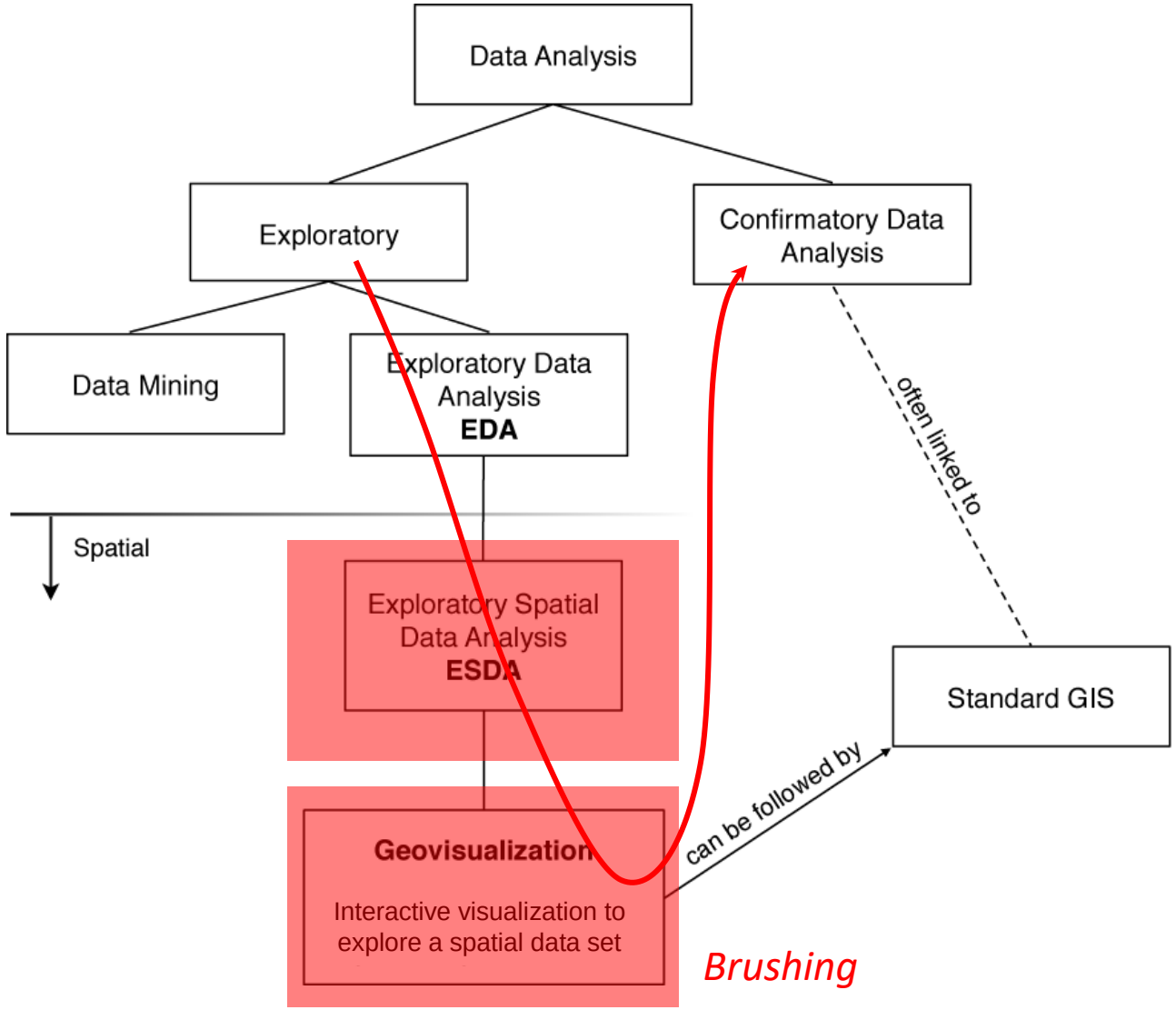
- Formulation of a single hypothesis
- To be verified by means of statistical tools (e.g. ANOVA)
- Focused investigation (only one goal)
- Works with theories and methods that are firmly established

Exploratory analysis



Exploratory spatial data analysis

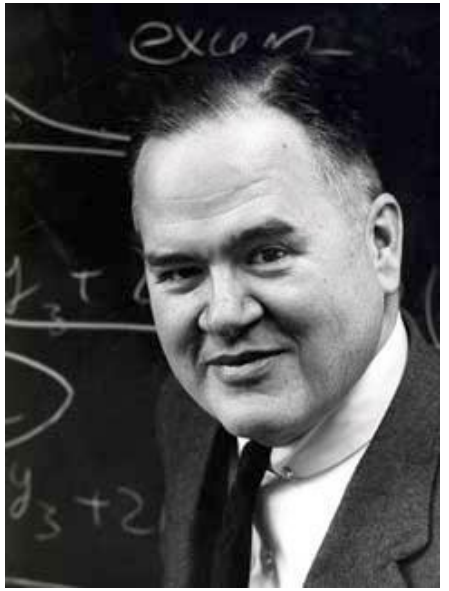
ESDA and/or **Geovisualization**



We need both exploratory and confirmatory

I assert, and I count upon most of you to agree after reflection, that to implement the very confirmatory paradigm (*) properly we need to do a lot of exploratory work.

Neither exploratory nor confirmatory is sufficient alone. To try to replace either by the other is madness. We need them both.



Tukey, J. W. (1980). We Need Both Exploratory and Confirmatory. The American Statistician, 34(1), 23–25. doi:10.2307/2682991

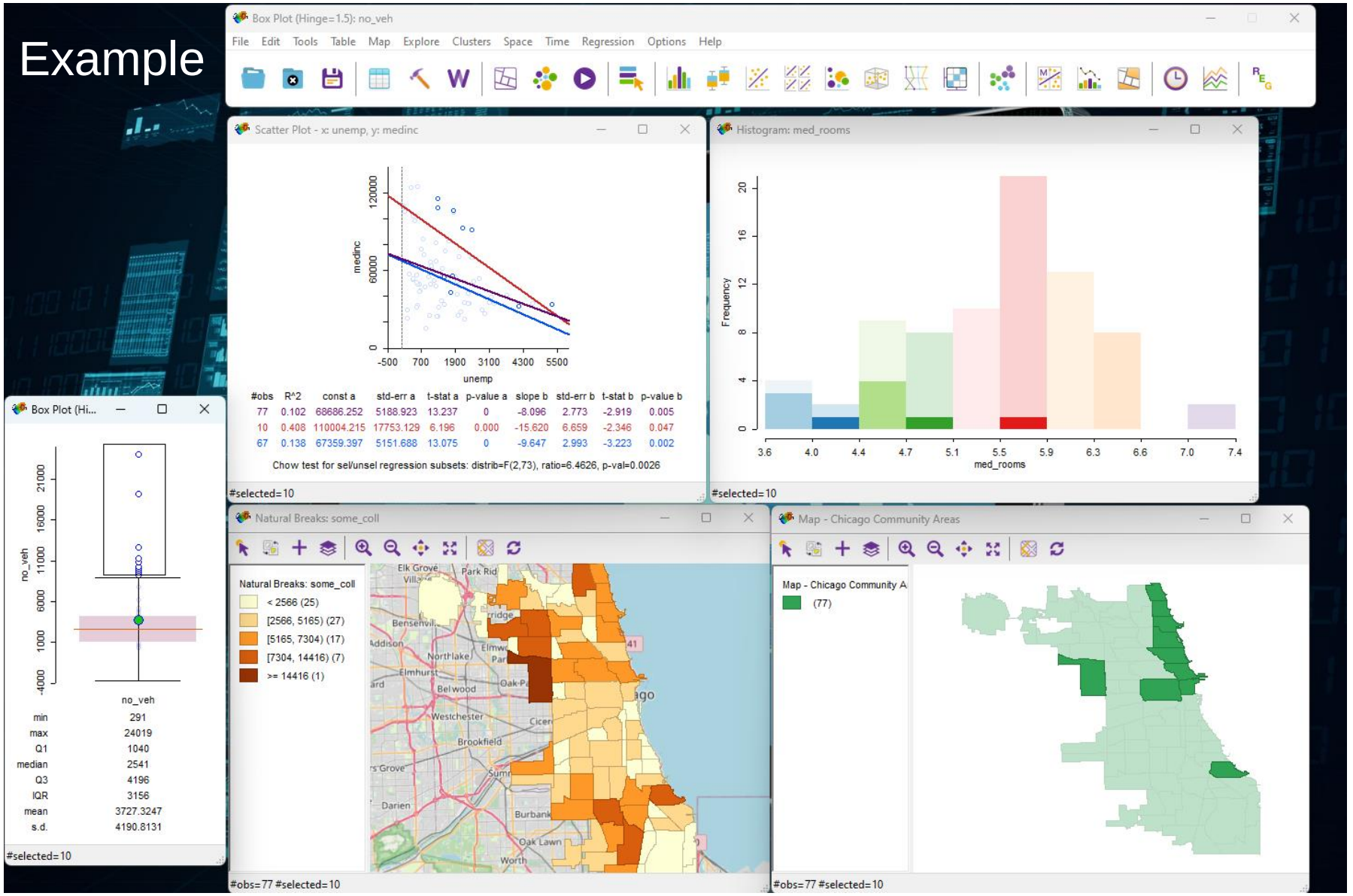
What is exploratory data analysis?

INTRODUCTION

An exploratory analysis looks at the data from as many angles as possible, always on the lookout for some interesting feature. The data analyst is interested in uncovering facts about the data and may use any procedure of his/her liking to this end. The only limits to such an analysis are those imposed by time constraints and the creativity of the data analyst. EDA is not guided by a desire to confirm the presence of a particular effect, and it is not supported by a statistical model that incorporates a mathematical expression for such an effect.

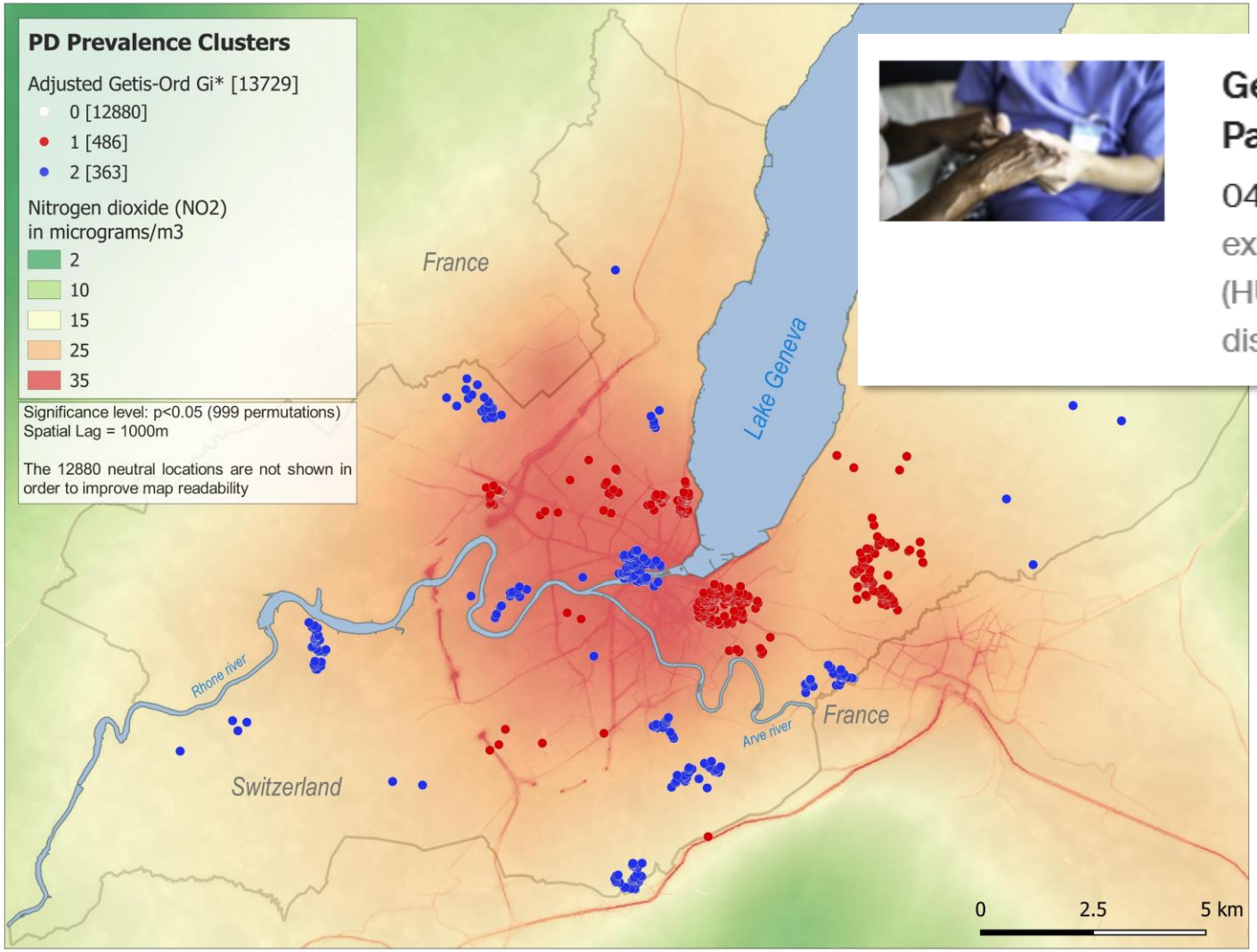
Morgenthaler, S. (2009). Exploratory data analysis. *WIREs Computational Statistics*, *1*(1), 33–44.

Example



Brushing

Application to environmental health (exploratory)



Geospatial data helps to better understand Parkinson's disease

04.02.21 — In a new paper, a team of EPFL spatial-analysis experts and neurologists from Geneva University Hospital (HUG) show that the probability of developing Parkinson's disease is higher in the canton of Geneva's urban centers...

Parkinson Register, Geneva

Application to environmental health (confirmatory)

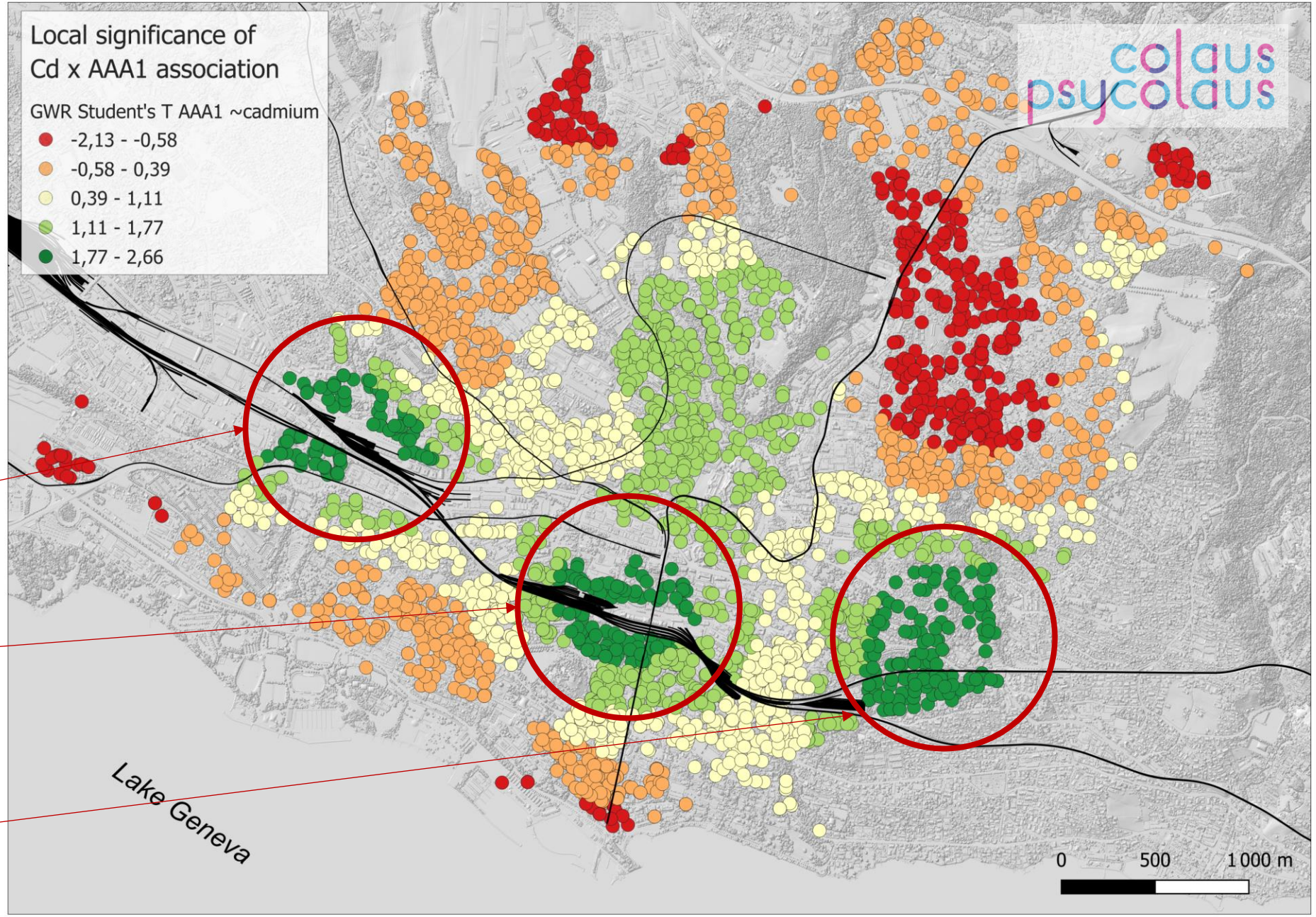
AAA1 autoantibody x cadmium in urine

Local Student's T (local significance) →

Sébeillon marshalling yard

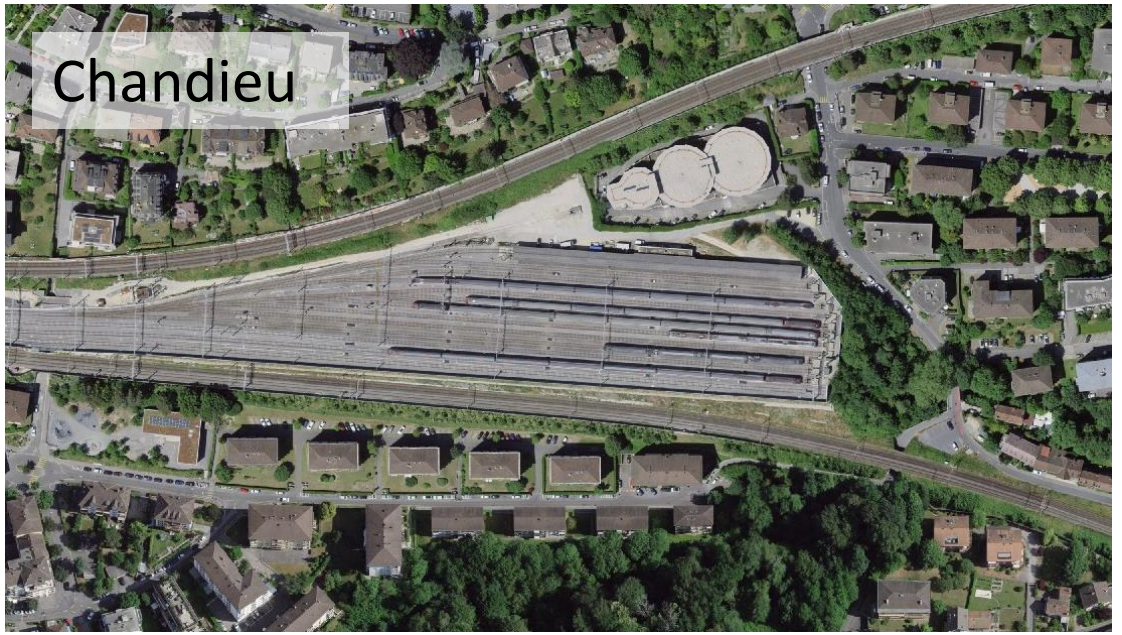
Lausanne main SBB station

Chandieu marshalling yard and train repair warehouses



- Potential sources of cadmium
- Friction with catenaries
- Brake systems
- Local industrial activities
- Altogether

- Rail transport increases cadmium levels in soil along railroad lines (Wilkomirski 2001; Ma et al. 2009)



Organization and requirements

Moodle: https://go.epfl.ch/edenv

Theoretical lectures

Components mentioned earlier are gradually addressed

1. What is exploratory data analysis (brushing), how to use this approach, statistical tools specific to EDA (order statistics, rate smoothing)
2. How can human cognitive skills be used in ESDA
3. Relationship between health and place (historic background)
4. Population health, population epidemiology (medical cohorts) and then spatial epidemiology
5. Specific exploratory approaches in ESDA: typology, classification, Principal Component Analysis (PCA)
6. Confirmatory statistics: spatial regression, geographically weighted regression

Exercises

- Exercise in the classrooms on your laptops (computer room also available)
- Open-Source software (Geoda, QGIS, R Studio)
- Exercises will be made available simultaneously with lectures
- Solutions will be made available on Moodle
- Short reports (compte-rendu), 1-2 pages free text to describe the exercise (sometimes to answer questions), 8/10 required: to be uploaded on Moodle
- **Deadlines for short reports**: assignment available each week with a deadline 11 days later, i.e. at the end of the next week, on Friday at 23h59.

ed Forum for questions

Semester project

- Collective work
- Composition of groups on week 5 (currently 9 groups of 5)
- Groups will have to produce:

1. A description of the semester project proposal (containing the research idea, the working hypotheses) - instructions will follow (Week 6)
2. An oral presentation of the research (last lecture of the class, Week 13)
3. A scientific article (~10 pages) as semester project (instructions about scientific paper writing on Week 6 also)

Data

- The main data (dependent variables) to elaborate the semester projects are the same for all groups:
  - 24 variables you will produce in the context of exercise 3 (open dataset of reference)
  - The health data - Body Mass Index (BMI) and frequency intake of sugar sweetened beverages (SSB)
- You can also download more explanatory data on the SITG website according to the research hypotheses developed (dependent variables are health data)
- The main open geodata source is the Système d'Information du Territoire Genevois (SITG)
- https://ge.ch/sitg/donnees/conditions-d-utilisation/open-data

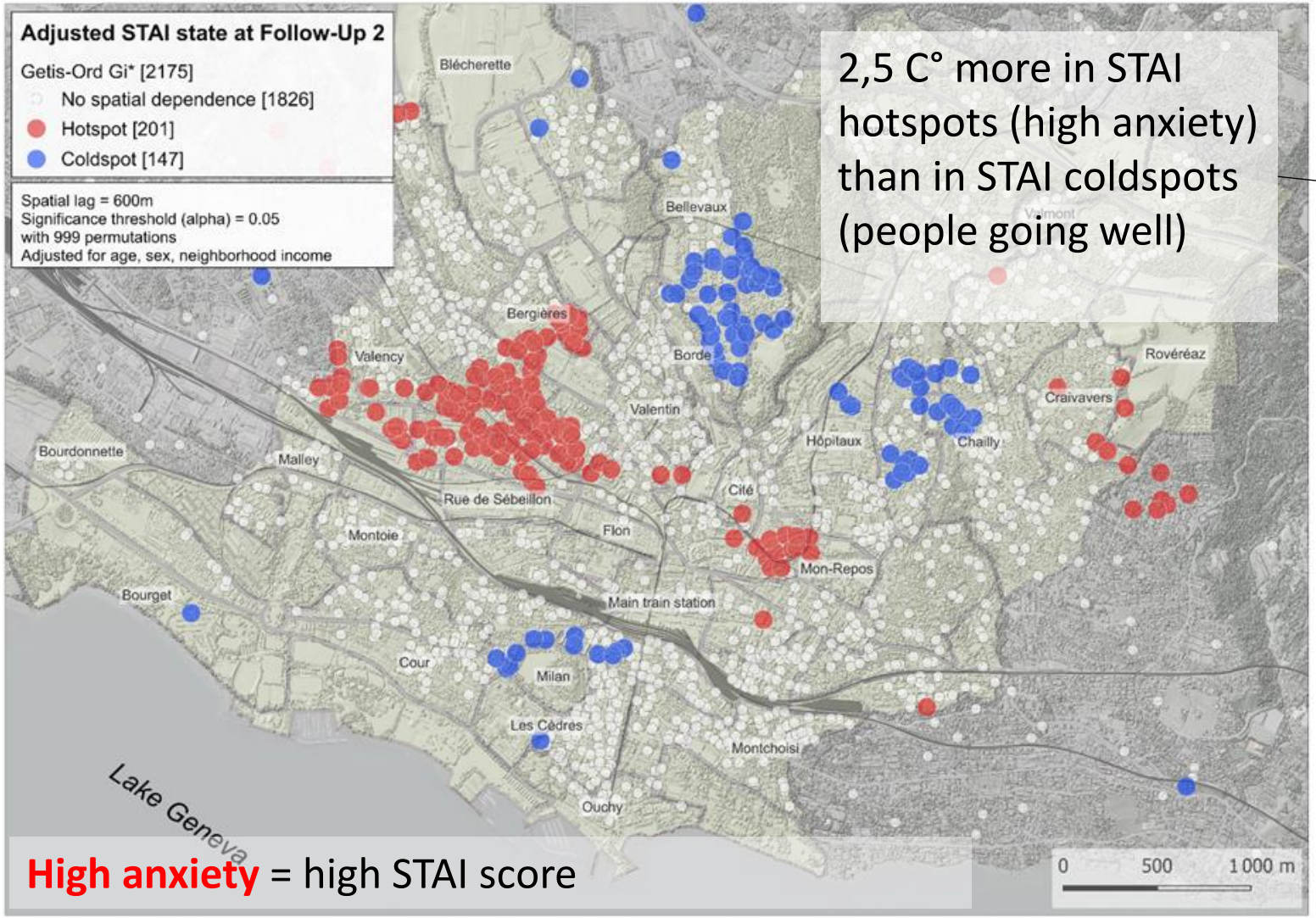
List of methods taught (or already known*)

- Order statistics (box plots, box maps, descriptive statistics)
- Confounding factors and variable adjustment
- Spatial statistics, spatial dependence (Global and Local Moran’s I*)
- Thematic mapping*
- Geographically Weighted Regression (GWR)
- Spatial regression
- Hierarchical Ascendant Classification (CAH)
- Principal Component Analysis (PCA)
- Spatial Relative Risk (SPARR)

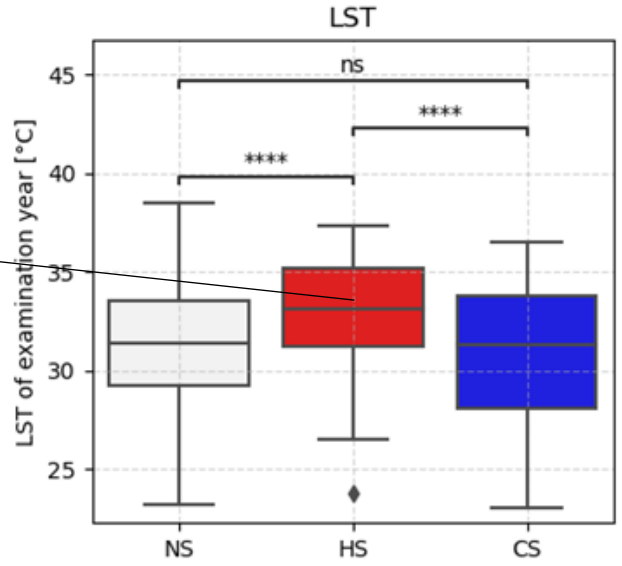
Spatial dependence of anxiety in Lausanne (2018-2021)

colaus psycolaus

Orrder statistics, Box plots



2,5 C° more in STAI hotspots (high anxiety) than in STAI coldspots (people going well)

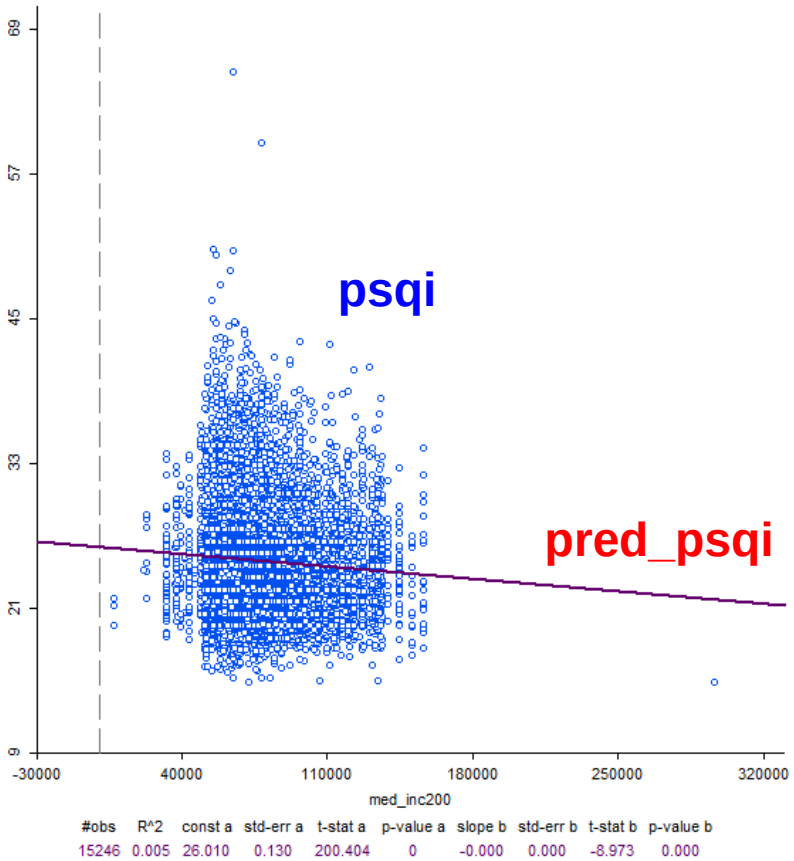


High anxiety = high STAI score

Spatial dependence (Getis-Ord Gi)

Sleep troubles in Geneva

Confounding factors and adjustment



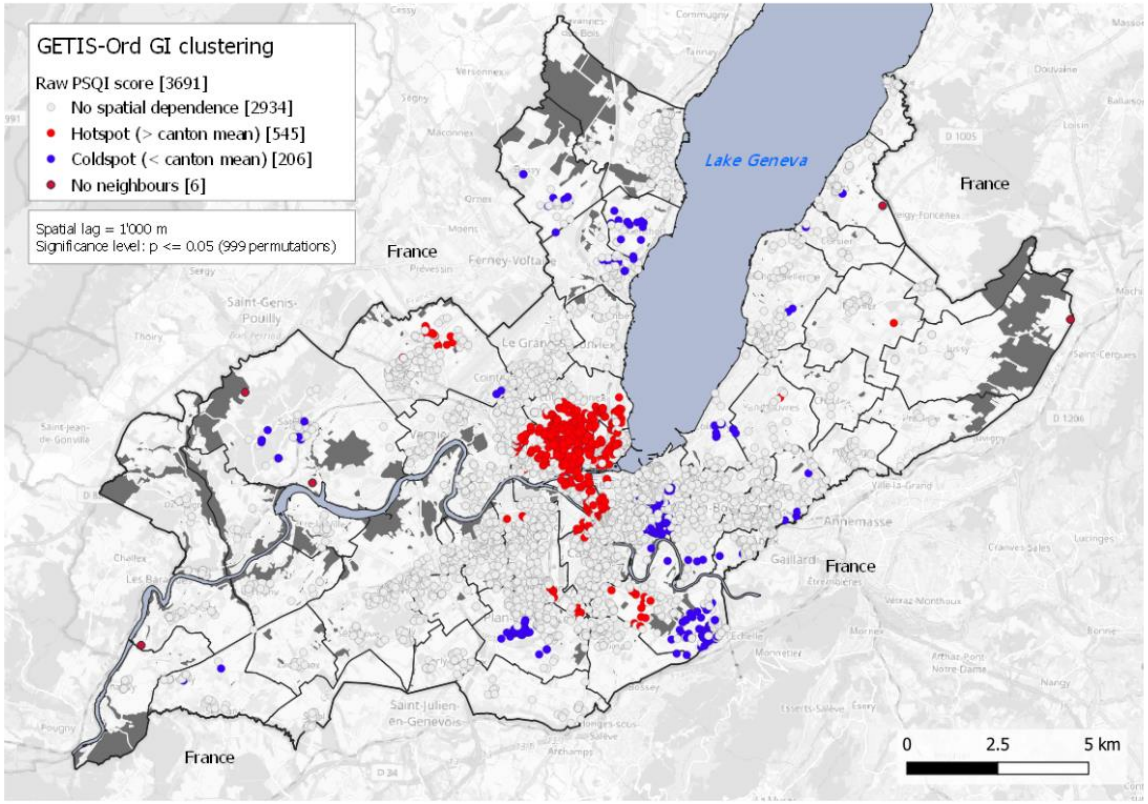
Income	psqi	pred_psqi	adj_psqi
45357	24.290592	26.0824925	23.4325795
40228	30.739416	26.34080248	29.62590952
73575	22.968384	24.66516142	23.52770258
65562	23.224413	25.06871111	23.38299789
56627	17.048376	25.51717464	16.76131336
43420	33.732502	26.17919629	32.78542821
43880	33.464035	26.15427369	32.53987331
50031	21.612812	25.84621886	20.99388914
56627	31.304895	25.51844091	31.01656609
56768	30.986309	25.51017701	30.70342799
44167	16.205019	26.1379469	15.2915521
56070	19.511112	25.54472585	19.19368215
39876	32.449974	26.35536755	31.32471845

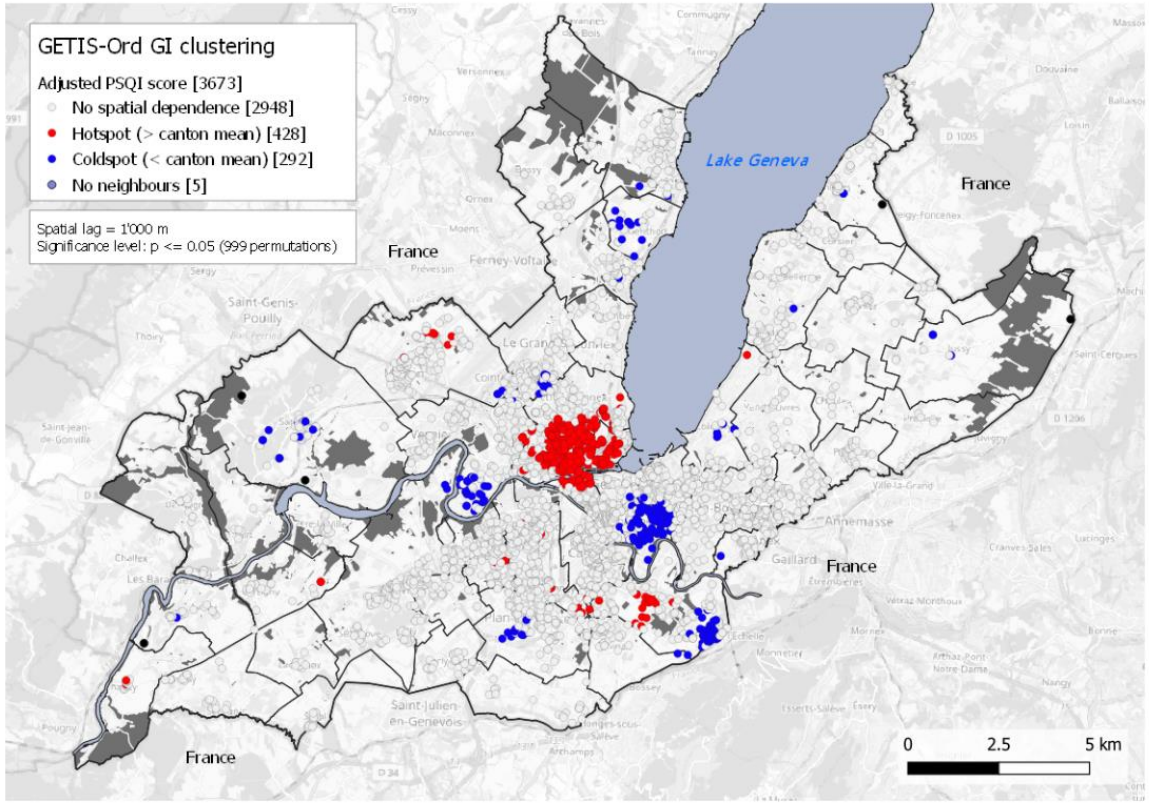
$$adj\_psqi = psqi - pred\_psqi + median(psqi)$$

SPECCHIO
Améliorons ensemble notre santé

Sleep troubles in Geneva

Confounding factors and adjustment

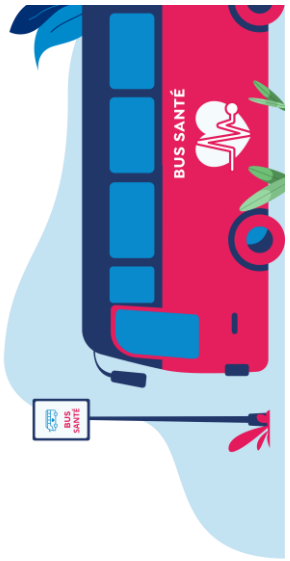


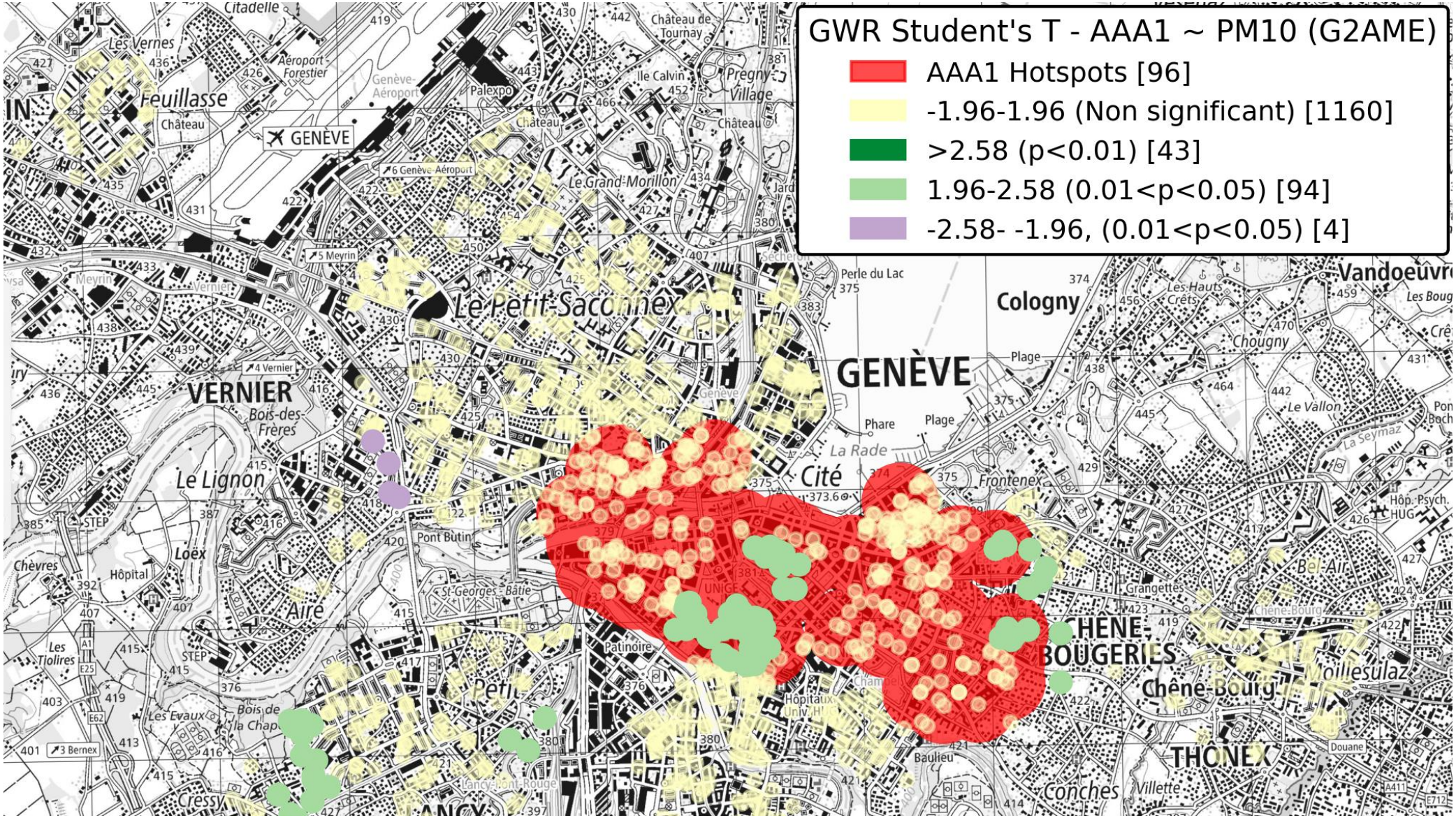


SPECCHIO
Améliorons ensemble notre santé

Local vs global relationship

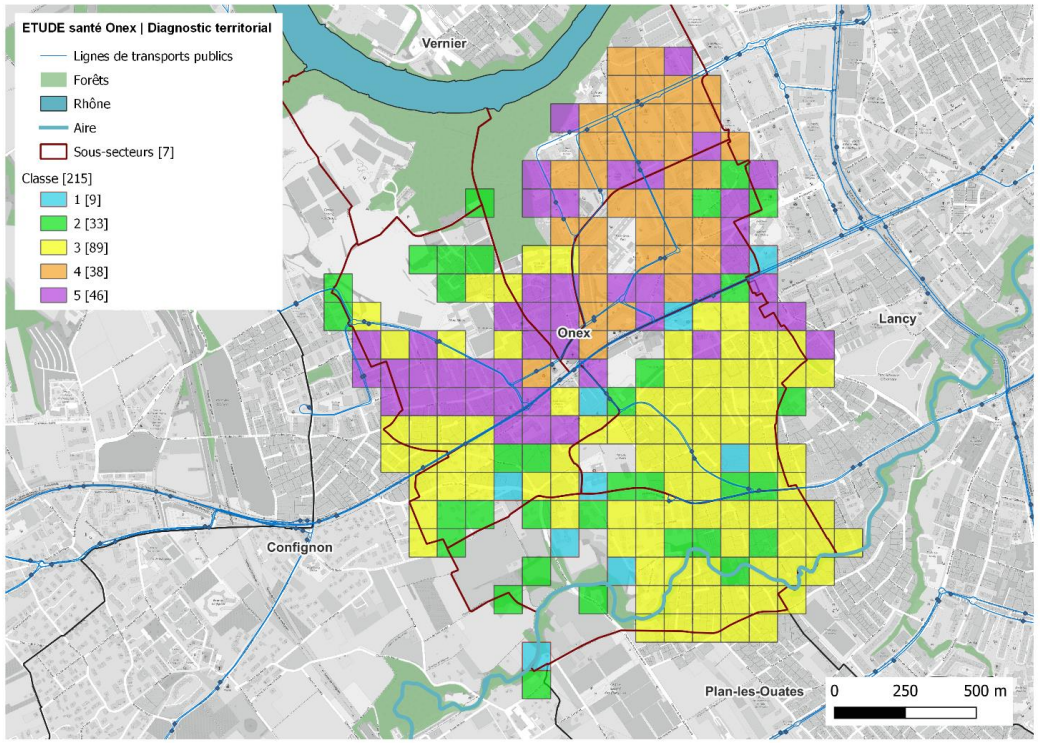
GWR





Health Territorial Diagnostic, Onex, Geneva

Spatial clustering, PCA



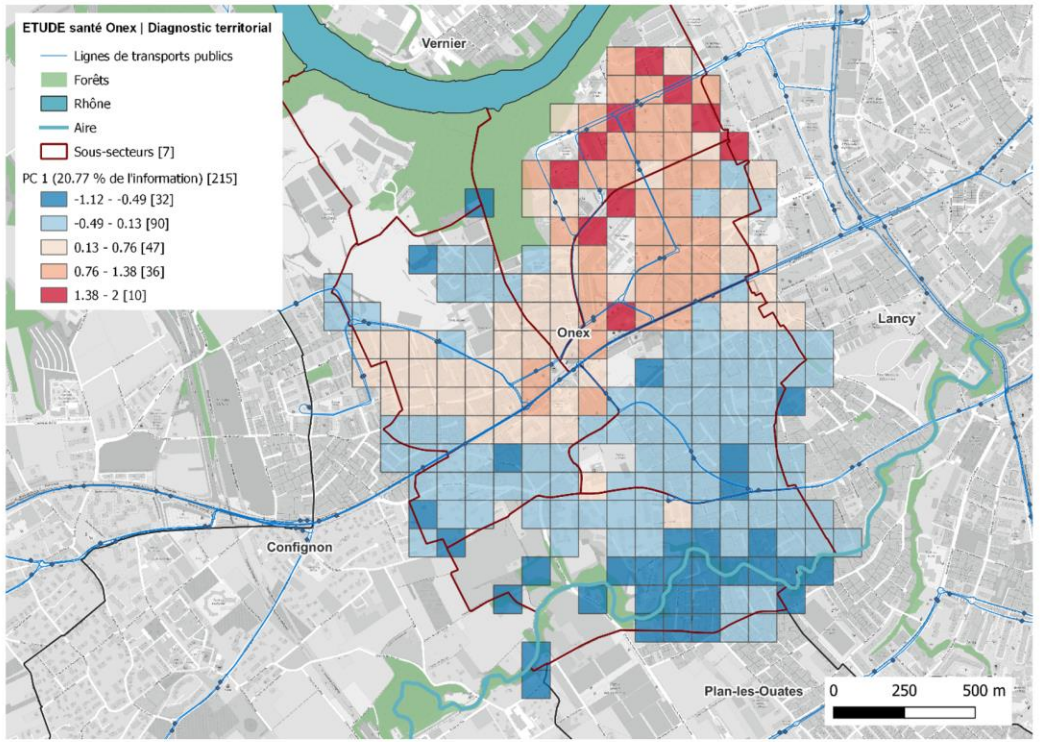
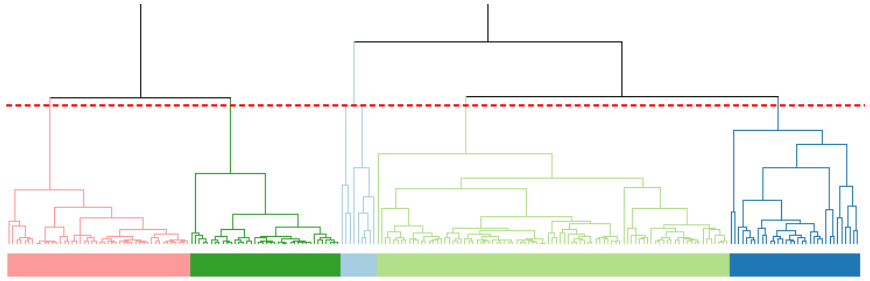
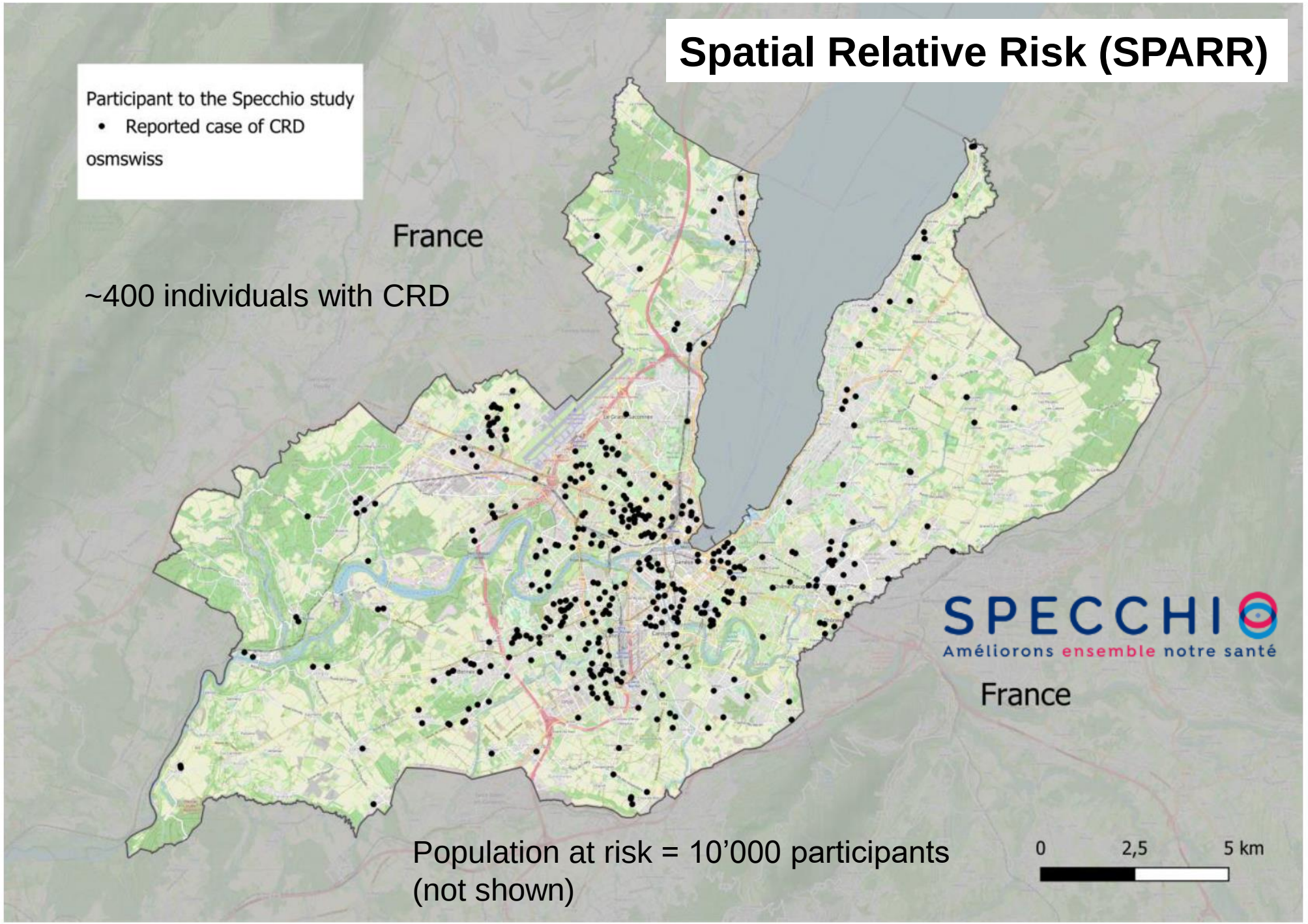


Figure 34 : Score factoriel par hectare sur la composante principale 1 (PC1). Ce score traduit l'intensité de la relation entre le comportement de la population dans les hectares et le profil traduit par la première composante. Le profil de cette dernière correspond majoritairement à une population dense de seniors allophones et socio-économiquement vulnérables. Le profil est décrit en détail dans le texte. Les couleurs sont similaires à celles utilisées sur la Table 4 (bleu = négatif, rouge = positif).

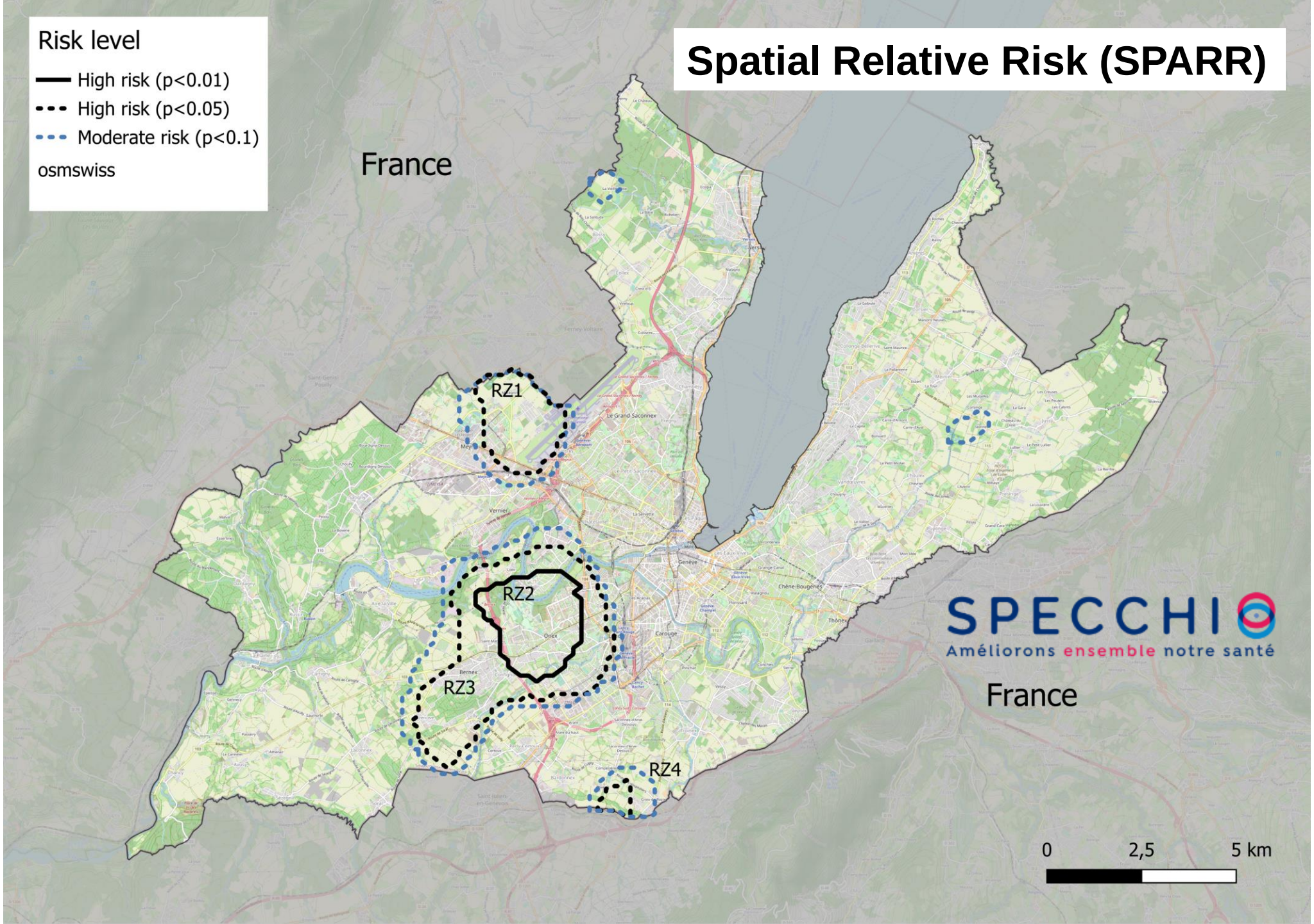


Relative risk

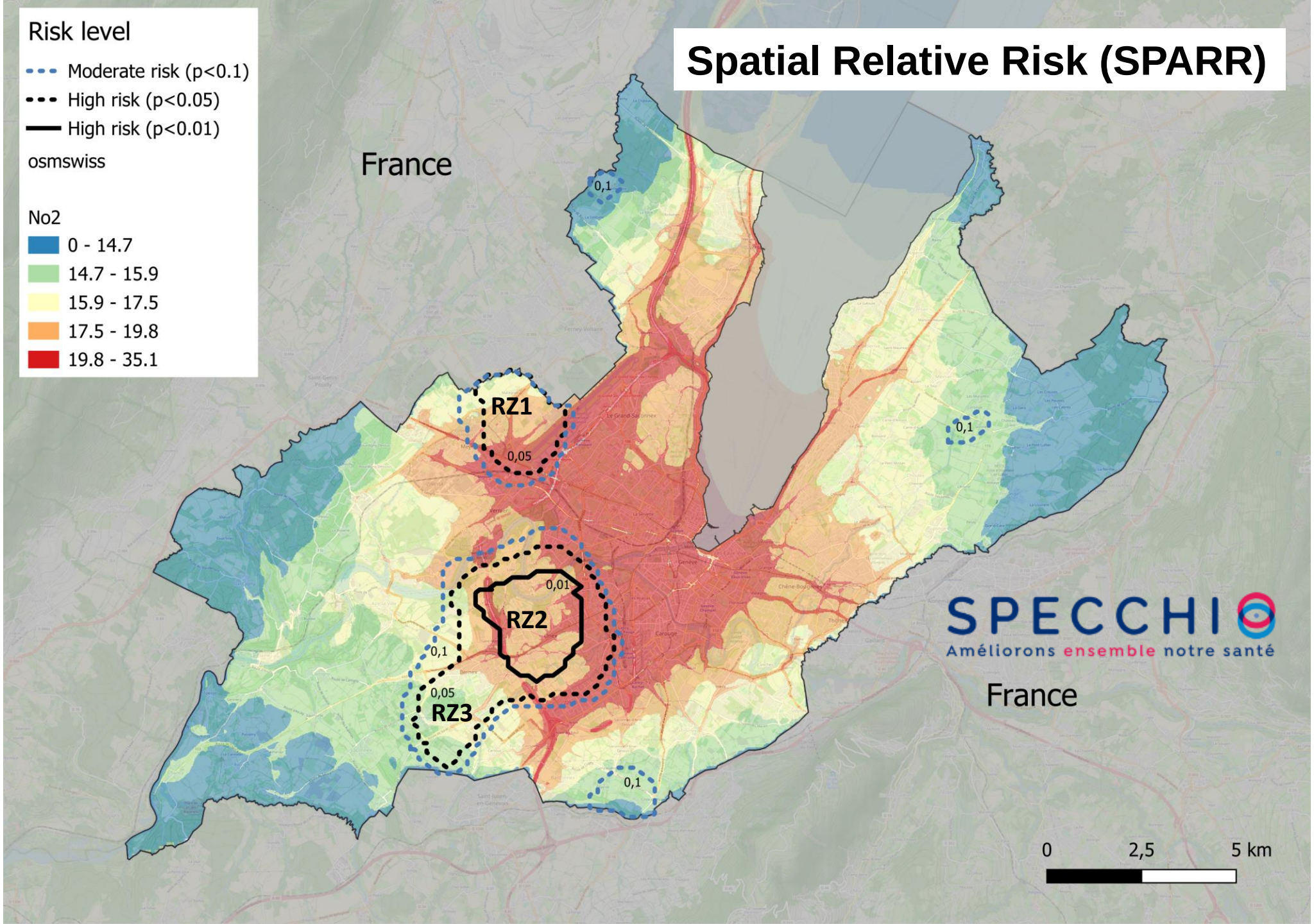
Spatial Relative Risk (SPARR)



Relative risk



Relative risk



Documentation

1. Geoda Workbook
2. Slides of the theoretical lectures
3. Scientific articles distributed during the course

Evaluation

- Continuous control during the semester and final oral exam

- Exercises (individual) = **20%** (8/10 short reports submitted and accepted)
  Short reports are not graded. I check that the work was correctly carried out. Report is ok (1) or not (0).
- A scientific article (semester project, group work, graded) = **30%**
- Oral exam (individual) = **50%**
  - Questions about the theoretical content of the course
  - Questions about the scientific paper written

Moodle

- The introduction contains the information presented here
- New versions of the programme may be made available in this introductory section [Programme of the course] v1 date
- Follow the ed Forum
- We have no TA: share your questions and answers on ed

Moodle: https://go.epfl.ch/edenv

Programme overview

EPFL | COURSEBOOKS

/ Study plans / Coursebooks / Exploratory data analysis in environmental health

ENV-444

Dr Stéphane Joost
Dr Mayssam Nehme

Theory

Exercises

Time for group work

	Period 1 **Monday 8^{15} à 9^{00}**	Period 2 **Monday 9^{15} à 10^{00}**	Period 3 **Monday 10^{15} à 11^{00}**	Period 4 **Monday 11^{15} à 12^{00}**
Week 1 **Introduction** September 9, 2024	**S. Joost** – Introduction to exploratory spatial data analysis and environmental health	**S. Joost** – Description of the course, explanations about requirements	**Exercise 1a** – Introductory readings (Morgenthaler and Anselin)	**Exercise 1b** – Introductory readings (Morgenthaler and Anselin) + short report writing = answer questions
Week 2 **Exploratory Spatial Data Analysis** September 23, 2024	**S. Joost** – Typical exploratory approach - Structuring spatial data analysis	**S. Joost** – Cognitive processes for geodata exploration	**Exercice 2a** – Basic data handling in Geoda	**Exercice 2b** – Histograms and other plots in Geoda + short report writing (include parts a and b)
Week 3 **Population epidemiology** September 30, 2024	**S. Joost** – Study of the relationship between health and place – The concept of exposome	**M. Nehme** – Introduction to population epidemiology	**Exercice 3** – Environmental dataset for Geneva	**Exercice 3** – Environmental dataset for Geneva + short report writing
Week 4 **Spatial epidemiology** October 7, 2024	**S. Joost** – Introduction to spatial epidemiology	**Exercice 4** – Health data handling and aggregation	**Exercice 4** – Health data handling and aggregation + short report writing	Start discussing the constitution of groups
Week 5 **Order stats and rate smoothing** October 14, 2024	**S. Joost** – Order statistics and rate smoothing	**S. Joost** – Confounding factors and variable adjustment	**Exercice 5** – Confounding factors and variable adjustment + short report writing	Constitution of ~8 groups Finalize group composition in Moodle
Holiday				

Programme overview

Week 6 **Writing of a scientific paper** October 28, 2024	**S. Joost** – Structure of a scientific paper – Collaborative writing and open data publication	**S. Joost** – Instructions for the description of the group semester project	**Exercice 6** – Prepare and upload open dataset to Zenodo	Time for group discussion and work on project
Week 7 **Medical cohort studies** November 4, 2024	**S. Joost** – Geographically Weighted Regressions (GWR)	**M. Nehme** – Medical cohorts, presentation of Specchio and Bus santé studies	**Exercice 7** – Geographically weighted Regression (GWR)	Submission of **description of the semester project** **Deadline**: Nov 8, 23h59
Week 8 **Spatial clustering** November 11, 204	**S. Joost** – Hierarchical Ascendant Classification (HAC) and Principal Component Analysis (PCA)	**S. Joost** – Exploratory Spatial Data Analysis for the analysis of cancer screening participation rate	**Exercise** 8 – PCA and HAC with GeoDa	Time for group discussion and work on project
Week 9 **Relative risk** November 18, 2024	**S. Joost** – Spatial Relative Risk (SPARR)	**Exercice 9** – Spatial Relative Risk (SPARR)	**Exercice 9** – Spatial Relative Risk (SPARR)	Time for group discussion and work on project
Week 10 **Metabolic syndrome** November 25, 204	**S. Joost** – Spatial regression	**Exercice 10** – Spatial regression	**M. Nehme** – Environmental pollution and metabolic syndrome	Time for group discussion and work on project
Week 11 **Thematic mapping** December 2, 2024	**S. Joost** – Thematic mapping, synthetic reminder	**S. Joost** – Analytical Design (how to improve thematic maps)	Time for group discussion and work on project	Time for group discussion and work on project
Week 12 **Work on semester project** December 9, 2024	Time for group discussion and work on project	Time for group discussion and work on project	Time for group discussion and work on project	Time for group discussion and work on project
Week 13 **Presentation semester projects** December 16, 2024	Presentation of semester projects (collective presentation)	Presentation of semester projects (collective presentation)	Presentation of semester projects (collective presentation)	Presentation of semester projects (collective presentation)
Deadline : submission of collective scientific paper on January 10 (Friday), 2025, at 23h59				

Why to write short reports ?
Why to write an article ?

“Writing helps you think because it gives you nowhere to hide”

“If you can’t clearly and simply define the words and terms you are using, you don’t really know what you’re talking about”

Richard Feynman

“I never teach my students; I only provide the conditions in which they can learn.”

Albert Einstein

Exercise 1 – Two readings

Morgenthaler

Overview

Exploratory data analysis

Stephan Morgenthaler*

Exploratory data analysis, or EDA for short, is a term coined by John W. Tukey for describing the act of *looking at data to see what it seems to say*. This article gives a description of some typical EDA procedures and discusses some of the principles of EDA. © 2009 John Wiley & Sons, Inc. *WIREs Comp Stat* 2009 1 33–44 DOI: 10.1002/wics.2

Keywords: descriptive statistics; box plot; median polish; smoothing with medians; *h*-distribution

INTRODUCTION

An exploratory analysis looks at the data from as many angles as possible, always on the lookout for some interesting feature. The data analyst is interested in uncovering facts about the data and may use any procedure of his/her liking to this end. The only limits to such an analysis are those imposed by time constraints and the creativity of the data analyst. EDA is not guided by a desire to confirm the presence of a particular effect, and it is not supported by a statistical model that incorporates a mathematical expression for such an effect.

With such a broad mandate, it is difficult to structure a presentation of EDA. We could follow Tukey's lead and use the type of data as a framework. In Tukey,[1] the first four chapters deal with a single series or multiple series of observations of the same variable. The next five chapters are about regression-like situations, and the next four deal with tables or arrays of observations in which the margins describe different circumstances. Further topics, in particular, questions related to counted data are treated in the last eight chapters. Another possible framework for discussing EDA procedures is the broad attitude underlying the method. These include the *classical mode* of thinking, which uses a specific model and derives procedures that are appropriate for that model, for example, those based on the likelihood function. A second broad attitude is the *exploratory mode*. Flexible, forgiveness, and ease of computation are the main characteristics of such procedures. As an illustration, Tukey used the image of a sailboat that can go anywhere, does not easily tip over, and is easy to sail. A third mode englobes *rough confirmatory* procedures, which are used to identify findings that merit a closer study. If a careful study is undertaken, methods that work well in the face of a variety of realistic circumstances are called for. Tukey called this *mustering strength*. In the remainder of this entry, we will make some general comments on the philosophical foundations and then focus on some procedures that are typical for EDA.

WHAT IS EDA?

When a statistician is asked to assess a series of observations $y_1, \ldots, y_n$ taken under identical conditions, he or she is most likely to compute the average $\bar{y} = \sum_{i=1}^{n} y_i/n$, the standard deviation $s^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1)$, a Student's t confidence interval for the mean, and maybe a χ^2-based confidence interval for the variance. When modeling the observations as independent realizations drawn from a normal population with unknown mean μ and variance σ^2, these procedures are fully justified. They can also be justified in a weaker, asymptotic or least squares or permutation sense. In most instances such models and their conditions are merely implicit and not explicitly specified.

In the EDA mode, on the other hand, there is no need to consider a model. The data are regarded as a list or batch of numbers, not necessarily representing an underlying population. As the name EDA suggests, one is free to choose any procedure to analyze the data, and the primary aims are to look at the data and to think about the data from many points of view. Informal conclusions are drawn in this manner. Graphical visualization is usually the first order of business. In the case of a single list of numbers, examples of such graphical displays include the drawing of a single axis with the observations indicated by some symbol (called a dot plot), a histogram, a normal quantile plot, or a box plot. These show the data in more or less detail. The choice of procedure may depend on what one has already learned about the data. If a histogram shows two distinctive modes or if the number of observations

*Correspondence to: stephan.morgenthaler@epfl.ch
Institute of Mathematics, Ecole polytechnique fédérale de Lausanne, Lausanne, Switzerland
DOI: 10.1002/wics.2

Anselin

Geographical Analysis ISSN 0016-7363

GeoDa: An Introduction to Spatial Data Analysis

Luc Anselin[1], Ibnu Syabri[2], Youngihn Kho[1]

[1]Spatial Analysis Laboratory, Department of Geography, University of Illinois, Urbana, IL, [2]Laboratory for Spatial Computing and Analysis, Department of Regional and City Planning, Institut Teknologi, Bandung, Indonesia

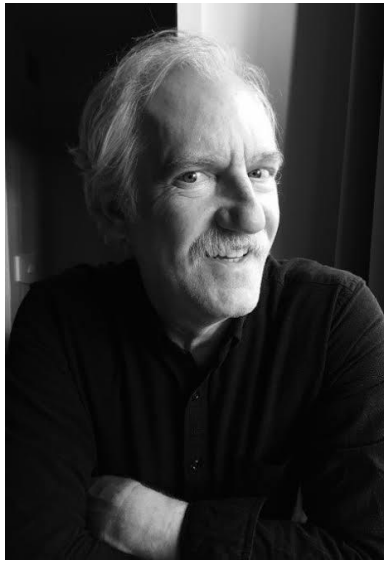
This article presents an overview of GeoDa™, *a free software program intended to serve as a user-friendly and graphical introduction to spatial analysis for non-geographic information systems (GIS) specialists. It includes functionality ranging from simple mapping to exploratory data analysis, the visualization of global and local spatial autocorrelation, and spatial regression. A key feature of* GeoDa *is an interactive environment that combines maps with statistical graphics, using the technology of dynamically linked windows. A brief review of the software design is given, as well as some illustrative examples that highlight distinctive features of the program in applications dealing with public health, economic development, real estate analysis, and criminology.*

Introduction

The development of specialized software for spatial data analysis has seen rapid growth as the lack of such tools was lamented in the late 1980s by Haining (1989) and cited as a major impediment to the adoption and use of spatial statistics by geographic information systems (GIS) researchers. Initially, attention tended to focus on conceptual issues, such as how to integrate spatial statistical methods and a GIS environment (loosely versus tightly coupled, embedded versus modular, etc.), and which techniques would be most fruitfully included in such a framework. Familiar reviews of these issues are represented in, among others, Anselin and Getis (1992); Goodchild et al. (1992); Fischer and Nijkamp (1993); Fotheringham and Rogerson (1993, 1994); Fischer, Scholten, and Unwin (1996); and Fischer and Getis (1997). Today, the situation is quite different, and a fairly substantial collection of spatial data analysis software is readily available, ranging from niche programs, customized scripts and extensions for commercial statistical and GIS packages, to a

Correspondence: Luc Anselin, Department of Geography, University of Illinois, Urbana-Champaign, Urbana, IL 61801
e-mail: anselin@uiuc.edu

Submitted: January 1, 2004. Revised version accepted: March 10, 2005.



Exercise 1 – First short report

- Answer the questions
- Submit your answers as a PDF on Moodle (assignment each week)
- No immediate deadline. I take short reports into account until Friday, September 20, at 23h59


Thank you for your attention!