UNIL | Université de Lausanne

Faculté des géosciences
et de l'environnement

# SPATIAL DATA SCIENCE 2020

## Book of Abstracts

## Local Organizing Committee

**Marj Tonini**, Dr. – Program committee & General Chair

**Mikhail Kanevski**, Prof. - Academic founder & Session Chair

**François Bavaud**, Prof. - Academic founder & Session Chair

**Federico Amato**, Dr. - Program committee & Session chair

**Christian Kaiser**, Dr. – Event Management Software manager

**Raphaël Ceré** - Website developer & Communication

**Raphaël Bubloz** - Event Management Software editor

**Romain Loup** - Event Management Software editor

**Fabian Guignard** - Abstract management

**Noé Carette** - Book Editing

### International Scientific Committee

Mariaelena Cama, GIS Specialist, Planner – Stuttgart (Germany)

Roberto Castello, Solar Energy and Building Physics Laboratory, EPFL, Lausanne (Switzerland)

Urska Demsar, School of Geography & Sustainable Development, University of St Andrews (Scotland, United Kingdom).

Stéphane Joost, Laboratory of Geographic Information Systems, EPFL, Lausanne (Switzerland)

Raphael Huser, King Abdullah University of Science and Technology (KAUST) (Saudi Arabia)

Gregoire Mariethoz, Institute of Earth Surface Dynamics, University of Lausanne (Switzerland)

Luigi Lombardo, Faculty of Geo-Information Science and Earth Observation, University of Twente (Netherlands)

Nahid Mohajeri, University of Oxford (United Kingdom) & EPFL Lausanne (Switzerland)

Beniamino Murgante, School of Engineering – University of Basilicata (Italy)

Thomas Opitz, Biostatistics and Spatial processes, INRA, Avignon (France)

Céline Rozenblat, Institute of Geography and Sustainability, University of Lausanne (Switzerland)

Sebastiano Trevisani, University Iuav of Venice (Italy)

Demyanov Vasily, Heriot-Watt University; School of Energy, Geoscience, Infrastructure and Society; Edinburgh (Scotland, United Kingdom)

Claudia Vitolo, European Centre for Medium-Range Weather Forecasts, Reading (United Kingdom).

Michele Volpi, Swiss Data Science Center, ETH Zurich / EPFL (Switzerland)

# Introduction

**Spatial Data Science 2020 (SDS2020)** is the first scientific meeting on this topic organized by scientists of the Faculty of Geosciences and Environment, University of Lausanne (Switzerland). It took place as virtual meeting in June 8-11, 2021.

The main objective is to initiate a dialogue about the different issues we face when performing and developing innovative methodologies of mining, analysis, modelling and visualization of geo-spatial data. The recent development of quantitative methods allowing to perform intelligent data reduction and suitable analysis is a central issue in environmental and socio-economic sciences. In both these fields, geo-referenced numerical data are nowadays massively available and can be further enhanced by other sources of information numerically transformed. Nevertheless, this information is often complex and sometimes unstructured or noisy. Thus, discovering interesting spatial or intrinsic patterns is a challenging task that led scientists to search for new tools.

With this in mind, innovative techniques based on clustering, pattern recognition, and data mining can be employed to extract knowledge and insights from data. In addition, new formalisms need to be developed to directly incorporate other sources of information in the characterization of the geographical space. These theoretical developments proved to be very helpful in various applications. Here just some examples: natural hazard susceptibility and risk assessment (e.g. flood, landslides, earthquakes, wildfires); multivariate time series analyses, for both environmental risks (e.g. pollution time series) and renewable energy potential assessments (e.g. meteorological data, such as wind speed, rainfall, solar radiation); understanding network flows (such as commuter traffic).

Nevertheless, several issues need to be addressed to improve these approaches, such as, for example, information bias and information noise, scale and mapping unit, selection of the predictor variables (redundancy/irrelevance and overfitting related problems), uncertainty.

The **main theoretical topics** of this workshop include, but are not limited, to three principal axes:

- Spatial quantitative methods using a strong statistical/mathematical framework, especially focused on the quality of formalism of the method;
- Geovisualization, with a major accent on visual analytics and computational data mining techniques focused on high-dimensional attributes;
- Pattern recognition and modelling, with special emphasis on approaches based on data-mining and machine learning.

The **main applications** are closely related to the research in environmental sciences, quantitative geography and spatial statistics, in particular:

- Natural and anthropogenic hazards (e.g. flood, landslides, earthquakes, wildfires, soil, water and air pollution);
- Socio-economic sciences, characterized by the spatial dimension of the data (e.g. census data, transport, commuter traffic).
- Spatio-Temporal Spread of Covid Patterns at local and global scale;
- Renewable energy resources spatio and/or temporal modeling and prediction (e.g. solar, wind, biomass, wave, geothermal energy).

June 2021
Lausanne, Switzerland Dr. Marj Tonini

# Table of content

KEYNOTE SPEECHES

ORAL PRESENTATIONS

**June 9th**

**June 10th**

**June 11th**

# KEYNOTE SPEECHES

## On Spatial Data Science

Prof. Mikhail Kanevski,
*Faculty of Geosciences and Environment, University of Lausanne (Switzerland)*

Data science is an emerging scientific discipline concerned with the development and application of theoretical and computational methods to work with and extract knowledge from Data. (Geo)statistics, geoinformatics and machine learning are basic and complementary methodological approaches contributing to the spatial data science (SDS). In the current presentation the main attention is paid to the analysis, modelling, prediction and visualisation of complex spatial (spatio-temporal) environmental data. A problem-oriented approach, which starts with the objectives of the study and quality and quantity of data, is adapted. It follows a generic data driven methodology: from data collection via intelligent exploratory data analysis and modelling with careful validation and testing to the interpretability/explainability of the results. SDS is considered as an experimental science, therefore experimentation with data by applying different methods, algorithms and tools is considered as very important. Such point of view helps in better understanding of data and phenomena, obtaining reliable and robust results and making intelligent decisions. The presentation is accompanied by simulated and real data case studies. In conclusion some general remarks and future perspectives are discussed.

**Biography:**

Mikhail Kanevski is Professor on Environmental data mining & Geostatistics. His current scientific interests cover a wide range of topics: geographical information science, environmental modelling, spatial statistics, time series forecasting, machine learning and environmental data mining. The major applications deal with natural hazards, environmental pollution and renewable energy analyses and assessments. He is. co-author of several books on spatial data modelling, introducing geostatistics and machine learning algorithms in the environmental sciences, namely: "*Analysis and modelling of spatial environmental data*" (2004), along with the Geostat Office software; "*Advanced Mapping of Spatial Data. Geostatistics, Machine learning, Maximum Bayesian Entropy*" (2008); "*Machine Learning for Spatial Environmental Data. Theory, Applications and Software*" (2009). The models developed and adapted by the group of Prof. Kanevski were successfully applied to geo-, environmental and socio-economic spatio-temporal data analyses. The fundamental scientific research was supported by several SNSF grants. At present Prof. Kanevski is a co-PI (collaboration with EPFL) of the PNR75 "Big Data" project "*Hybrid renewable energy potential for the built environment using big data: forecasting and uncertainty estimation*".

**Practical strategies for fitting Extreme-Value statistical models with a view towards environmental and ecological applications**

Dr. Daniela Castro Camilo,
*Lecturer in Statistics, University of Glasgow, UK*

Over the last years, Extreme-Value Statistics (EVS) has gained considerable attention in environmental science, as extreme observations have increased in size and frequency. Mathematically, EVS has a well-developed asymptotic framework that allows us to study extreme events of single or multiple processes observed in one or many locations over space and time. Moreover, it enables us to make statements regarding future events that can be even more extremes than those observed. The mathematical elegance of these methods faces a couple of challenges in the applied arena. For instance, most asymptotically justified EVS models are computationally expensive, and their application to spatial data is limited to few locations. Moreover, some of these models cannot account for well-known features in environmental data, such as decaying dependence strength as events become more extreme. Other problems are related to constraints imposed by the limiting models that do not naturally exist in the observed processes.

In this talk, I will present three different approaches to tackle the previous issues. The first approach is a computationally appealing method to model multiple extreme events over spatially rich regions that successfully captures weakening extremal dependence. The second and third approaches leverage the integrated nested Laplace approximation (INLA) framework, which allows fast and accurate inference in complex models applied to data with different levels of spatial coverage. We will see how to apply these methodologies using precipitation, wind speed, fishery and pollution data. I will conclude with some reflections on how EVS can be incorporated into widely used classical statistical models.

**Biography:**

Daniela Castro-Camilo is a Lecturer in Statistics at the University of Glasgow. Her research focuses on the theory and applications of multivariate and spatial extremes, with a particular interest in environmental, geological, and ecological applications. During the last few years, her work has gravitated around the integrated nested Laplace approximation (INLA) method for Bayesian inference. Specifically, she has developed methods promoting the need to adequately capturing extremes observations within the usual statistical analysis centred around mean values. She has worked closely with INLA developers to implement and improve extreme value models to help to bridge the gap between statistical theory and practice. She co-authored the book "Advanced spatial modelling with stochastic partial differential equations using R and INLA" (CRC Press, 2018).

### Understanding the spatial structure of cities: some results and challenges

Prof. Marc Barthelemy,
*Institut de Physique Théorique, CEA, CNRS-URA, France*

The recent availability of large amounts of data about cities allowed us to better understand the spatial organization of cities and how they evolve in time. In this talk I will present a small selection of results and also discuss some challenges. I will first present some tools for the characterization of infrastructure networks (such as roads and subways) and their temporal evolution. I will then discuss mobility patterns obtained from mobile phone data and the polycentric structure of cities. If time allows, I will end this talk by discussing theoretical and empirical challenges about urban sprawl.

**Biography:**

Marc Barthelemy is a former student of the Ecole Normale Supérieure of Paris. In 1992, he graduated at the University of Paris VI with a thesis in theoretical physics titled "Random walks in random media". Since 1992, he has held a permanent position at the CEA (Paris) and since 2009 is a research director at the Institute of Theoretical Physics (IPhT) in Saclay and a member of the Center of Social Analysis and Mathematics (CAMS) at the Ecole des Hautes Etudes en Sciences Sociales (EHESS). He has worked on applications of statistical physics to complex networks, epidemiology, and more recently on spatial networks. Focusing on both data analysis and modeling, he is currently working on urban networks and various aspects of the emerging science of cities. Marc Barthelemy co-authored the book "Dynamical Processes on Complex Networks" (Cambridge Univ. Press., 2008), and published recently the books "The Structure and Dynamics of Cities" (Cambridge Univ. Press, 2016) and "Morphogenesis of spatial networks" (Springer 2018).

**Interactive deep learning for animal conservation from above**

Prof. Devis Tuia,
*Associate Professor, Environmental Computational Science and Earth Observation Laboratory, Ecole Polytechnique Fédérale de Lausanne (EPFL)*

Monitoring wildlife populations is a complex business, since it involves monitoring over large areas, with complex terrains and counting living animals that move (and can also be dangerous at close range). For all these reasons, as well as to increase frequency and reduce costs, Unmanned Aerial Vehicles (UAVs) are more and more used. UAVs indeed acquire large amounts of data, but then also raise the problem of detecting and counting the animals, in order to provide accurate counts, in an automatic way. In this talk, I will talk about how deep learning can help, especially when helped by enthusiastic nature lovers willing to screen images for protecting wildlife.

**Biography:**

Devis Tuia received a Ph.D. in Environmental Sciences at University of Lausanne in 2009. He was then a postdoc researcher at the University of València, Spain, the University of Colorado, Boulder, CO, USA and EPFL Lausanne. In 2014-2017, he was an Assistant Professor at the University of Zurich. He is now full professor at Wageningen University, the Netherlands. Since 2020, I joined EPFL Valais, to start the ECEO lab, working at the interface between Earth observation, machine learning and environmental sciences. His research focuses on geospatial computer vision, a field at the interface between GIscience, remote sensing and machine learning. He develops digital solutions to address problems of land planning and the environment. He led most of his efforts in urban recognition, land-use modeling and analysis, but he also has experience in wildlife tracking, environmental risk reduction and forest management through scientific collaborations.

**Current works:**

(i) Making remote sensing accessible to everyone ! Developing algorithms for human machine interaction; (ii) Open the black box: interpretable deep learning and uncertainties in environmental modeling; (iii) Digital wildlife conservation: using imaging to automatize censuses and conservation efforts.

# ORAL PRESENTATIONS

## The need to account for spatial landslide data bias effects in statistically-based landslide modelling

Steger S.[1], Mair V.[2], Kofler C.[1], Pittore M.[1], Zebisch M.[1], Schneiderbauer S.[1],
*Eurac Research, Institute for Earth Observation, Bolzano-Bozen, Italy[1], Office for Geology and Building Materials Testing, Autonomous Province of Bolzano-South Tyrol, Cardano-Kardaun, Italy[2]*

Supervised classification algorithms are frequently applied to identify landslide-prone areas (i.e. landslide susceptibility maps) or to get insights into the static causes of slope instability. However, the landslide inventory data used to train the underlying models is often affected by a systematic spatial incompleteness (e.g. underrepresentation of movements in woodlands or in remote areas). Thus, the quantity to be modelled (i.e. the response variable) may not perfectly represent the spatial distribution of the phenomena of interest. Literature reveals that the effects of such landslide data biases are often ignored when interpreting data-driven landslide susceptibility models.

This research was built on the basis of landslide data from the province of South Tyrol (7400 km$^2$) that systematically represents damage-causing events and ignores landslides far from infrastructure. The created models (M1, M2, M3) represent diverse strategies to handle spatially biased landslide data. The goals were to show why geomorphic cause-effect relationships cannot always be deduced from models that exhibit an apparent high predictive performance (M1), to evaluate the usefulness of a bias-correction approach under serious data bias conditions (M2) and to exploit the underlying data bias to map areas affected by damage causing landslides (M3). The models were critically evaluated by means of statistical associations, variable importance ranking, performance and plausibility.

The presented research may offer an alternative perspective on how flaws in available landslide information can be considered in data-driven landslide modelling. It is demonstrated that under common landslide data bias conditions, the focus should not only lie on the actual geomorphic process (landslide susceptibility effects), but also on the respective landslide data context (landslide data collection effects). The findings showed that none of the three models was able to create a useful representation of landslide susceptibility, despite calculated high predictive performances. In most cases, geomorphic causation could not be deduced by interpreting the modelled relationships between landslide inventory data and the environmental factors. The final impact-oriented model (M3) enabled us to identify (temporally independent) damaging landslides with high accuracy. We conclude that despite the availability of increasingly flexible models and automated variable selection techniques, a thorough qualitative investigation of landslide data limitations will remain essential towards meaningful spatial landslide models. An inference of geomorphic causation may be challenging under landslide data bias conditions, even though model performance indicators might suggest a high model quality.

## Detection of landslide clusters in Italy using space-time scan statistics

Pecoraro G.[2], Tonini M.[1], Romailler K.[1], Calvello M.[2],
*University of Lausanne[1], University of Salerno[2]*

Landslides are very common natural hazards in regions with relief and mountains. These events can result in severe consequences to humans, in terms of dead, injured and evacuated people. In addition, they can cause environmental and infrastructural damages with significant economic losses. Historical landslides inventories represent a precious source of information for risk assessment, allowing to investigate the pattern distribution and temporal evolution of these events. The acquisition and analysis of historic data of landslide events, although comprising only basic information, is essential for evaluating and managing landslide risk at small scales. The temporal dependency among persistent events is of paramount importance in landslide susceptibility, hazard and risk assessment. However, ordinary maps reporting the density of landslides recorded in a given area (often based on the computation of "landslide indexes") only consider the spatial dimension, neglecting its interaction with the time of occurrence of the events. Spatio-temporal Scan Statistics is the perfect tool to overcome this issue, since it allows detecting statistically significant excess of observations (i.e. clusters) thanks to moving windows that scan all locations both in space and in time.

The present study addresses the pattern distribution of recent landslides in Italy. The main objective is the detection and mapping of spatio-temporal clusters of landslides that occurred in the period 2010-2017 in the country. To this aim, a subdivision of the national area into 158 warning zones, as identified by the 21 civil protection regional centers to deal with weather-induced hydro-geological hazards, is adopted. Information on landslides comes from FraneItalia, a georeferenced catalog developed consulting online news sources. Analyses are performed both at national scale and at a regional scale, focusing on the Campania region. The space-time permutation scan statistics model is applied to detect statistically significant clustering, accounting for the geographical spatial dimension and for the temporal dimension. Two types of analyses are performed: annual, considering each single year; and multi-annual, encompassing the entire 8-year study period. In both cases, spatio-temporal cluster analyses are able to detect areas and frame-periods characterized by relevant and recurrent landslide activity. Finally, obtained results are compared with a standard landslide density map, highlighting the complementarity of the two approaches.

### References:

Tonini M., Pecoraro G., Romailler K., Calvello M., 2020. Spatio-temporal cluster analysis of recent Italian landslides. Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards. https://doi.org/10.1080/17499518.2020.1861634

## A pattern recognition approach for strong following earthquake forecasting

Gentili S.[2], Di Givambattista R.[1],
*Istituto Nazionale di Geofisica e Vulcanologia, Italy[1], National Institute of Oceanography and Applied Geophysics - OGS, Italy[2]*

Numerous strong earthquakes are followed by one or more subsequent large earthquakes (SLE), of similar magnitude or even greater than the initial earthquake. Repeating earthquakes cause accumulated damage to already weakened buildings and infrastructures; therefore, forecasting their occurrence is a challenging task from the viewpoint of civil protection, to stop the continuous loss of lives and to reduce economic damages. Studies concentrate on the value of Dm, defined as the difference in magnitude between the mainshock and the strongest SLE. We classify the earthquake clusters into two classes: Dm≥1 (swarms or large aftershock seismic sequences) as type A, otherwise (smaller aftershocks seismic sequences) as type B. To forecast if the cluster following a strong earthquake will be a type A, we proposed a decision tree-based pattern recognition approach, which uses seismological features extracted from the first hours/days seismicity of the cluster. In particular, we analyzed features based on the space and time distribution and energy of earthquakes in the cluster. The method, called NESTORE (Next STrOng Related Earthquake), has been successfully applied to Northeastern Italy and Western Slovenia (Gentili and Di Giovambattista, 2020) and to Italy as a whole (Gentili and Di Giovambattista, 2017). In this study, we will present the results of a new improved version of NESTORE in which we refine the screen out procedure for the less informative features. NESTORE method is adaptive and depends on the analyzed region; during the supervised training phase, some features are selected as the best-performing ones in the analyzed area and are subsequently being used for classification. Rigorous statistical tests of this method will be proposed for Southern California seismicity, based on SCSN earthquake catalog. We analyzed all clusters with mainshock magnitude ≥4 from 1981 to today. The method shows a percentage of A and B clusters correctly classified up to over 80%. Some tests have been performed on Northern California and Western Nevada clusters extracted from the Comprehensive Earthquake Catalog, by using NESTORE trained using Southern California clusters. The good performances suggest that the seismicity characteristics are similar in the two regions.

This research is partially supported by the Italian Ministry of Foreign Affairs and International Cooperation for a bilateral project Italy-Japan entitled “Analysis of seismic sequences for strong aftershock forecasting”.

## Distribution of monogenetic volcanism along the Cameroon Line

Schmidt C.[2], Laag C.[3], Profe J.[1],
*Justus-Liebig-University Gießen[1], University of Lausanne[2], Université de Paris[3]*

Volcanic eruptions may constitute a severe threat for local communities and their infrastructure. Important information as the prediction of future eruption sites and the likelihood of activity can be obtained by analysis of spatio-temporal eruption pattern in an area of interest. The fact that monogenetic volcanoes, unlike polygenetic ones, erupt only once (within a geologically short period) at a certain spot and then volcanic activity jumps to another spot, renders a quantitative, probabilistic assessment of eruptive cycles challenging. In other words, the purely temporal risk assessment relevant for polygenetic volcanism has to be coupled with the spatial dimension in case of monogenetic volcanic fields to allow for a combined spatio-temporal forecast.

While the eruption history of many stratovolcanoes along the Cameroon Line (CL) in Central Africa is relatively well studied, only fragmentary data exists on the distribution and timing of monogenetic volcanism (mainly scoria cones and maars), presumably associated with Quaternary timescales. Here, we undertake an initial step in filling this gap and present for the first time a map of monogenetic volcanic features for most parts of the CL. Scoria cones and maars were identified by their characteristic morphologies using a combination of field knowledge, digital elevation models and satellite imagery. More than ~1100 scoria cones and 55 maars were identified and divided into eight monogenetic volcanic fields (MVF), as defined by the convex hull of the outermost vents: Bioko, Mt. Cameroon, Kumba, Tombel Graben (including Mt. Manengouba), Noun, Oku, Adamawa, and Biu (Nigeria). However, due to the rugged topography in the Oku volcanic field and the difficulty of identifying volcanic features remotely, the number of mapped scoria cones appears rather incomplete.

While the delineation of individual MVF contains an inherent subjective moment, statistical analyses of the primary dataset clearly shows that the mean nearest neighbor distance increases from <1 km to ~2 km from the oceanic sector (Bioko, Mt. Cameroon) in the southwest towards the continental part in the northeast (Adamawa, Biu). Correspondingly, the areal density of monogenetic features decreases along this gradient by about one order of magnitude from >0.2 km$^{-2}$ (southwest) to 0.02 km$^{-2}$ (northeast). This finding is in general agreement with prior geochronological results, indicating increased Quaternary activity towards the central and oceanic part of the CL (e.g., Njome and de Wit, 2014). Testing spatial organization of monogenetic volcanoes by the Geological Image Analysis Software (GIAS, v2; Beggan and Hamilton, 2010) reveal that vents are clustered (98% credible interval) in all MVF. Therefore, conclusions can be drawn on the tectonic control of (future) eruption locations. Finally, in order to provide for the first time an objective hazard estimate for monogenetic volcanic eruptions along the CL, each volcanic field is investigated by kernel density estimates (KDE) for spatial vent density. This is realized with the R-package ‘ks’ (Duong et al., 2021) based on the unconstrained ‘SAMSE’ bandwidth selector (Duong and Hazelton, 2003).

**References:**

Beggan, C., Hamilton, C.W., 2010. New image processing software for analyzing object size-frequency distributions, geometry, orientation, and spatial distribution. Computers & Geosciences 36, 539-549.

Duong, T., Hazelton, M., 2003. Plug-in bandwidth matrices for bivariate kernel density estimation. Journal of Nonparametric Statistics 15, 17-30.

Duong, T., Wand, M., Chacon, J., Gramacki, A., 2021. ks: Kernel Smoothing. R-package version 1.12.0. https://CRAN.R-project.org/package=ks

Njome, M.S., de Wit, M.J., 2014. The Cameroon Line: Analysis of an intraplate magmatic province transecting both oceanic and continental lithospheres: Constraints, controversies and models. Earth-Science Reviews 139, 168-194.

## A diachronic mapping of soil organic content in Croatia

Trevisani S.[1], Bogunovic I.[2],
*University IUAV of Venice[1], University of Zagreb, Faculty of Agriculture, Department of General Agronomy, Zagreb, Croatia[2]*

The spatiotemporal analysis of soil properties in areas of intensive agriculture is crucial for multiple reasons, including the monitoring of soil fertility, nutrient deficiency, the evaluation of anthropic impacts, and the planning of restoration/preservation practices. However, the comparison of the current soil fertility with the older situation is a difficult task, having to cope with the frequent problem of unbalance in sampling density between current and past conditions. In this context, the present case study, related to the diachronic analysis of soil organic content in Slavonia and Baranja region (Croatia), is emblematic of a typical example of data unbalance. The objective is to compare current spatial distribution of soil organic matter (OM) content with the setting monitored in the 60's of the past century. The accuracy of soil organic content evaluation for past and present datasets is comparable; the same can be said for the sampling methodologies adopted. The sampling spatial densities of the two datasets are extremely different: the current data set is almost exhaustive (19386 samples) and historical data are extremely sparse (152 samples). Hence the data from the 60's are less than the 1% of current data set; from this perspective this case study is emblematic of a strong "data imbalance" from the perspective of spatial sampling densities. Another interesting characteristic is related to the wide spatial coverage of the study area, characterized by different morphologies, soil types, and lithology (Velić and Vlahović, 2009; Bašić, 2013; Bogunovic et al., 2017; 2018). Moreover, all the samples are collected from croplands. The case study presented has also other points of interests. It offers the possibility to study the evolution of soil OM content in a highly anthropized environment, characterized by a millennial history of agriculture practices and with a sharp increase in last 70 years. This kind of study is a novelty for Croatia, and it can be relevant for other European countries in the context of studying the agricultural impacts of soil exploitation. The key objective of the study is to compare the present spatial distribution of soil OM content with the one observed in the 60's of the past century.

### Acknowledgments:

This work was supported by Croatian Science Foundation through the project "Soil erosion and degradation in Croatia" (UIP-2017-05-7834) (SEDCRO).

### References:

Bašic´, F. (2013). The soils of Croatia. Dordrecht Heidelberg New York London: Springer.
Bogunovic, I., Trevisani, S., Seput, M., Juzbasic, D., Durdevic, B. (2017). Short-range and regional

spatial variability of soil chemical properties in an agro-ecosystem in eastern Croatia. Catena, 154, 50-62.

Bogunovic, I., Trevisani, S., Pereira, P., Vukadinovic, V. (2018). Mapping soil organic matter in the Baranja region (Croatia): Geological and anthropic forcing parameters. Science of the total environment, 643, 335-345.

Velic´, I., Vlahovic´, I. (2009). Geological map of the Republic of Croatia 1: 300.000. Croatian geological survey, Zagreb.

## Monte-Carlo Kriging: An application to insurance data

Rongiéras L.[1], Chautru E.[1],
*Centre de Géoscience, MINES ParisTech, Université PSL[1]*

The estimation of a random function at a spatial location can be given by Kriging methods, which aim to build a linear interpolation of the available data. In practice, the observations may suffer from a misplacement of their true location. A deep interest is taken here in the case where they are mistakenly assigned to a single point in space. In that case, classic Kriging methods are not applicable. We propose to adapt them to this particular context, using the scheme of Monte Carlo approaches. Considering all locations as random, it consists in first simulating several of their realizations, then computing the Kriging weights for each of these simulations and finally averaging them. Under several assumptions, a theorem is provided insuring the almost sure convergence of the empirical estimation. A study with simulated data illustrates the efficiency of the estimation of the variogram based on this method. Precisely, the induced location error creates a disturbance in the estimation of the variogram, for which several transformations can be applied to correct it. Two variograms estimates are studied. The first one is computed with located simulation. In practice, it has a tendency of overestimating the nugget effect although the sill is quite accurate. The second one is computed by averaging the random function realization for same points localization and then estimating the variogram. It usually provides the right nugget effect although the sill is underestimated. Furthermore, it can be in practice less stable than the first method, since fewer points are considerated. The transformation can be obtained using those variogram estimations. Then, the Monte Carlo Kriging is applied to insurance data. In particular, it is used to firstly identify the spatial correlation of the surface of apartments in the city of Lyon. Several indexes are used to understand the quality of the application of Monte-Carlo Kriging Secondly, it is applied in order to suggest an estimation of the surface of apartments. The impact of the different parameters is also presented. Indeed, the area of location simulation, the number of simulations, and the variogram picked to represent the spatial correlation are the main parameters of the model. Choosing the optimal set up can be challenging.

## Functional Peaks-over-threshold Analysis and its Applications

de Fondeville R.[2], Davison A.[1],
*EPFL[1], Swiss Data Science Center[2]*

Estimating the risk of natural hazards has become important in recent decades, but up until now it has been largely limited to re-using catalogs of historical events, which usually do not exceed 40 to 50 years in length, and to numerical models, which require heavy computation and are often unreliable for extrapolation.

Extreme Value Theory provides a theoretical framework to describe and model tails of statistical distributions within which estimating the frequency of past extreme events as well as to extrapolating beyond observed severities is possible. These have been extensively studied in a univariate framework especially for independent identically distributed replicates, and applications have been developed in numerous fields. Due to recent extreme events, there has been a surge of interest in environmental applications, motivated by the necessity to better understand the impact of global warming. Floods, windstorms, heatwaves have a complex spatio-temporal structure that cannot be modelled using univariate extreme value theory.

We present an extension of peaks-over-threshold analysis to functions which allows one to define complex extreme events as special types of exceedances, and then obtain their limit distribution for increasingly high thresholds, namely the generalized r-Pareto process. We focus on a specific model based on log-Gaussian random functions using classical covariance structures to characterize extremal dependence.

Finally, we describe a stochastic weather generator for extreme events, capable of quantifying the recurrence of past events as well as generating completely new ones. The methodology is illustrated with two applications: first, we present a spatio-temporal model to quantify the risk associated with European winter storms. Secondly, we study heavy rainfall events in the region of Zurich under different notions of risk in order to generate water level scenarios for the Sihl river whose water stream flows under the Swiss biggest train station.

## Estimating high-resolution Red Sea surface temperature hotspots, using a low-rank semiparametric spatial model

Hazra A.[1], Huser R.[1],
*King Abdullah University of Science and Technology (KAUST)[1]*

In this work, we build and fit a complex spatial statistical model to estimate extreme sea surface temperature (SST) hotspots, i.e. high threshold exceedance regions, for the Red Sea, a vital region of high biodiversity. Sea surface temperature has indeed an immense environmental and ecological impact on marine life and ecosystems, e.g. affecting the survival of endangered animal species including corals. It also has an important economic impact for neighboring countries, which depend on it for their local fisheries and tourism. Here we analyze high-resolution satellite-derived SST data comprising daily measurements at 16703 grid cells across the Red Sea over the period 1985-2015. To this aim, we propose a semiparametric Bayesian spatial mixed-effects linear model with a flexible mean structure to capture spatially-varying trend and seasonality, while the residual spatial variability is modeled through a Dirichlet process mixture (DPM) of low-rank spatial Student-t processes (LTPs). By specifying cluster-specific parameters for each LTP mixture component, the bulk of the SST residuals influence tail inference and hotspot estimation only moderately. Our proposed model has therefore a nonstationary mean, flexible covariance, tail dependence structures and also allows to perform efficient posterior inference through Gibbs sampling. In our application, we show that the proposed method outperforms some natural parametric and semiparametric alternatives. Moreover, we show how hotspots can be identified and we estimate extreme SST hotspots for the whole Red Sea, projected until the year 2100, by incorporating knowledge from climate model outputs based on the Representative Concentration Pathways 4.5 and 8.5. The estimated 95% credible region for joint high threshold exceedances includes large areas covering major endangered coral reefs in the southern Red Sea. This study confirms that mitigation measures are necessary to safeguard coral reefs and ecosystems in this region.

## Local indicators of multivariate spatial autocorrelation: a weighted formalism

Bavaud F.[1], Loup R.[1], Guex G.[1],
*University of Lausanne[1]*

This contribution exposes a general formalism aimed to tackle multivariate autocorrelation in a weighted setting, with two geographical illustrations.

Spatial structure can be specified by a distribution over pairs of regions: the joint probability to select two regions is a measure of their spatial interaction and defines the edge weights of the associated network. Vertex weights obtain as margins of the edge weights, and define the regional weights, reflecting their importance (population, area, wealth, etc.). Edge weights are not necessarily symmetric, that is the network can be oriented, but the row and column margins must coincide (marginal homogeneity). Row-standardized edge weights are the spatial weights entering in spatial auto-regressive models and constitute Markov transition matrices whose stationary distribution provides the regional weights.

Multivariate regional features, numerical or categorical, can be flexibly incorporated into the formalism as squared Euclidean dissimilarities, whose weighted inertia measures the global dispersion of the configuration [1]. Using the pair probability introduced above, rather than its expected value under independence, one obtains the local inertia instead of the global inertia. Comparing local and global inertias permits to define a multivariate Moran index, measuring spatial autocorrelation, whose statistical significance can be assessed through a modified modes permutation test, taking into account regional weights [2]. In this formalism, the distinction between Geary and Moran multivariate indexes becomes immaterial.

Also, spatially lagged quantities, local covariances and partial covariances turn out to enjoy simple formal properties. In particular, the diagonal components of the lagged scalar product between multivariate features happens to provide a natural definition of the local measure of multivariate spatial autocorrelation (LISA) [3][4]. Also, the (weighted) regression slopes of the Moran-Anselin univariate scatterplots exactly coincide with the corresponding (weighted) univariate Moran measures of spatial autocorrelation.
Two cases studies illustrate the formalism: (1) spatial autocorrelation of Guerry data (1833) "Moral Statistics of France" [5]; and (2) spatial autocorrelation of textual descriptions of tourist images (Flickr) of Lavaux (Switzerland) [6].

### References:

[1] Bavaud, F., Kordi, M., and Kaiser, C. (2018). Flow autocorrelation: a dyadic approach. The Annals

of Regional Science, 61(1), 95-111

[2] Bavaud, F. (2013). Testing spatial autocorrelation in weighted networks: the modes permutation test. J. Geogr. Syst. 3(15), 233-247

[3] Anselin, L. (1995). Local indicators of spatial association - LISA. Geographical analysis, 27(2), 93-115

[4] Anselin, L. (2018). A Local Indicator of Multivariate Spatial Association: Extending Geary's c. Geographical Analysis.

[5] Dray, S. and Jombart, T. (2011). Revisiting Guerry’s data: Introducing spatial constraints in multivariate analysis. Ann. Appl. Stat. 5 (2011), no. 4, 2278-2299

[6] Ceré, R., Kaiser, C., and Reynard, E. (2019). http://lavaux.unil.ch/

## Sentinel-1 SAR Level-2 OCN Offshore Wind Speed Time-Series Simulation using Multiple-Point Statistics

Hadjipetrou S.[1], Mariethoz G.[1],
*University of Lausanne*[1]

Offshore wind exploitation has brought favorable opportunities regarding renewable energy production and a promising future for energy systems. Assessing the offshore wind resource at the local or regional scales, however, has proved challenging. Remote Sensing data have been widely exploited in the literature to derive high spatial resolution wind fields and their variations. Weibull distributions are then typically fitted to the data time-series to derive the power density output via the distribution parameters. Sentinel-1A/B satellites span the globe providing wind speed estimates at 10m above the sea surface, at a spatial resolution of 1km. The repeat frequency of these satellites (6 days from 2016 onwards), however, limits the estimation of power density as it does not provide enough data (i.e. a few instances within a month's period).

In this study, we employ the Multiple-Point (Geo)Statistical framework to simulate wind speed images at 6-hour intervals, around the offshore area of Cyprus, thus filling the gaps of the Sentinel time-series both in space and time. Prior to the simulation, Sentinel-1 data have been validated against in-situ measurements from 3 coastal meteorological stations. Sentinel-1 validated data are subsequently used as Training Images (TI) along with Uncertainties in Ensembles of Regional Reanalyses (UERRA) data as auxiliary variable, after the latter had been downscaled at Sentinel's spatial resolution via cubic interpolation. The MPS algorithm used, Quick Sampling (QS), is proved to be computationally efficient while being able to reproduce spatial complex patterns along with their variability. As an illustration of the methodology, offshore wind speed images are simulated at a spatial resolution of 1km over a 1-year period. The results imply that MPS simulations could form a consistent time series of high spatial and temporal resolution wind speed images, leading to improved estimates of offshore wind power assessment for the region.

## Optimized spatial surveys of peat thickness utilising remote sensing data

Marchant B.[1],
*British Geological Survey*[1]

Peatlands are important habitats and store large quantities of carbon. If the UK is to meet its commitments to net zero carbon emissions, it is vital that the quality of peatlands is monitored and protected. Spatial surveys of peat quality based on ground measurements are costly, time consuming and laborious. We explore how more efficient surveys of peat thickness in Dartmoor can be conducted by utilising radiometric measurements from the Tellus survey (http://www.tellusgb.ac.uk/) of the southwest England. Peats attenuate the radiometric signal emanating from the underlying rocks and therefore the amplitude of the observed radiometric signal is inversely related to peat thickness. However, this relationship is complex, with the bedrock radiometric signal varying according to the geological setting, the rate of attenuation of this signal varying according to the moisture content and porosity of the peat, and the uncertainty in this relationship being largest for thick peats which absorb most of the signal. Therefore, some ground measurements are required to calibrate a locally applicable model. We extend standard geostatistical models to accommodate this complex relationship by including spline basis functions in the fixed effects and by permitting heteroscedasticity. The fixed and random effects and the degree of heteroscedasticity in this model are estimated simultaneously by maximum likelihood, and the smoothness of the spline basis function is selected automatically according to the resultant value of the Akaike Information Criterion. We use a spatial simulated annealing algorithm to optimize the spatial configuration of the required ground-based measurements of peat thickness in order to minimize the predicted average uncertainty in the spatial thickness predicted by this model. This predicted uncertainty function uses a Taylor series approximation to quantify the errors resulting from the model estimation in addition to the uncertainty resulting from interpolation of the measured peat thicknesses. When this approach is applied to data from Dartmoor, it leads to ground measurements being focused where the radiometric signals are small and peats thickness is expected to be large. A survey incorporating radiometric data and 30 ground measurements is found to be as precise as a survey of 300 measurements in the absence of radiometric data.

## Considering spatio-temporal dependence in k-fold cross-validation

Wang Y.[1],
*University of Twente, ITC[1]*

### Background:

Machine learning has been widely and successfully used in various geographical modelling researches and the built models' performances on existed samples should be evaluated to assess their practical prediction ability in actual application. K-fold cross-validation is the most commonly used method to evaluate built models. And random approach is usually adopted in k-fold cross-validation. However, random approach will lead to overoptimistic results in evaluations. These overoptimistic results are caused by spatial dependence, which is composed by spatial autocorrelation and spatial heterogeneity.

Currently, there are several k-fold cross-validation approaches that have considered spatial autocorrelation, such as buffer approach and block approach, or heterogeneity, such as spatial k-means approach, to overcome the random approach's overoptimistic evaluation results.

However, there are some shortcomings existing in current approaches. For example, buffer approach will lead to information loss then build poor predictive models in cross-validations. Thus, the non-overoptimistic evaluation result of buffer approach is "false negative". In block approach, samples in adjacent blocks might be all close to the same border, making the spatial autocorrelation hard to avoid. In spatial k-means approach, only spatial locations of samples are considered, but spatial heterogeneity not only includes the difference of locations but also contains the difference of environmental variables. In addition, current approaches also lack considering spatial autocorrelation and heterogeneity together. It should be improved in this research too.

At present, several studies have compared different approaches, but most of them have ignored an important issue that is setting a benchmark to claim which approach can produce a more accurate evaluation result. Except for samples used in cross-validation, that means part of samples should be selected as test samples, which are used to acquire model's prediction ability benchmark while they are also used to simulate model's actual application situation. Due to these various situations, in this research, the test samples also have different settings to correspond to various situations.

### Methodology:

Here we propose a new k-fold cross-validation approach to overcome the previous approaches' shortcomings mentioned above. This proposed approach contains two steps.

The first step is used to consider spatial autocorrelation. In first step, nearest-neighbor hierarchical clustering (NNHC) with maximum linkage is used to split samples into different blocks. Comparing with regular block approach, NNHC always merge the closest samples into the same block. Thus, samples in NNHC blocks are better to avoid spatial autocorrelation. NNHC block's size is in line with the regular block method, which is spatial autocorrelation threshold.

The second step is used to consider spatial heterogeneity. It divides NNHC blocks acquired above into k-folds by k-means clustering. But this k-means clustering is not only based on samples' locations but also based on environmental and target variables.

**Experiments:**

There are two case studies used in this research. The first one is Meuse River region's zinc case study, it contains 152 samples and covers around 11 km$^2$. The second one is California house-price case study, it contains 20640 samples and covers around 4.2 * 105 km$^2$.

There are 2 basic series of experiments designed: heterogeneity-majored experiments and autocorrelation-majored experiments. In heterogeneity-majored experiments, three kinds of test samples settings are implemented on both two case studies. These three settings are corresponding to three situations between model's-built region and applying region: different, partially different, similar. In autocorrelation-majored experiments, test samples settings also have three kinds, but are only implemented on California case study, because samples amount in Meuse River case study cannot support this series of autocorrelation-majored experiments. The three settings are high autocorrelated, medium autocorrelated, low autocorrelated.

In all experiments, 7 approaches are implemented. Except for proposed approach, there are 6 compared approaches: random, buffer, block, spatial k-means, proposed step 1 and proposed step 2. And each approach in each experiment is implemented 5 times to reduce random effects.

**Results:**

After experiments implementation, a relative MAPE of every approach will be calculated by this approach's MAPE (MAPE acquired by cross-validation) minus benchmark (test set MAPE). This relative MAPE is able to indicate if the corresponding approach is accurate (small relative MAPE value) or not (large value).

The results of the experiments proved that the proposed approach is able to more accurately evaluate model's prediction ability when this model is applied to somewhat different regions, or applying regions are not highly autocorrelated with regions that were used to build prediction models. By considering spatial autocorrelation and heterogeneity together, the proposed approach has achieved research objective that is to make models' evaluation results acquired by k-fold cross-validation more reliable, not overoptimistic and not over-negative.

**References:**

Brenning, A. Spatial prediction models for landslide hazards: review, comparison and evaluation. Natural Hazards and Earth System Science 2005, 5, 853–862.

Gasch, C.K.; Hengl, T.; Gräler, B.; Meyer, H.; Magney, T.S.; Brown, D.J. Spatio-temporal interpolation of soil water, temperature, and electrical conductivity in 3D + T: The Cook Agronomy Farm data set. Spatial Statistics 2015, 14, 70–90.

Belgiu, M.; Dragut, L. Random Forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing 2016, 114, 24–31.

Broennimann, O.; Guisan, A. Predicting current and future biological invasions: Both native and invaded ranges matter. Biology Letters 2008, 4, 585–589.

Oliveira, M.; Torgo, L.; Santos Costa, V. Evaluation Procedures for Forecasting with Spatio-Temporal Data; Springer, Cham, 2019; pp. 703–718.

Bahn, V.; McGill, B.J. Testing the predictive performance of distribution models. Oikos 2013, 122, 321–331.

Schratz, P.; Muenchow, J.; Iturritxa, E.; Richter, J.; Brenning, A. Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecological Modelling 2019, 406, 109–120.

## Building simulation models coupling territorial and network dynamics at the interface of disciplines and scales

Raimbault J.[1],
*University College London[1]*

Interactions between transportation networks and the dynamics of land-use are crucial to take into account when planning and managing sustainable urban environments. A quantitative understanding of such processes using models has been proposed by several disciplines, including Land-use Transport Interaction models or economic models of transport network growth. A co-evolution approach (in terms of circular causal relations) to modelling interactions between transportation networks and territories can be proposed, to simulate the dynamics of territories on long time scales. Such a viewpoint has however not been extensively explored, as the question is by nature interdisciplinary and at the interface of spatial and temporal scales.

The purpose of this communication is to synthesise results obtained with simple agent-based and simulation models at different scales and integrating paradigms from different disciplines, from planning to transport geography, urban geography, physics and economics. These models have the common feature of strongly coupling territorial and transportation network dynamics.

A first model at the scale of the urban area explores the coupling between a population density generation model based on aggregation-diffusion processes with multiple network growth heuristics. This co-evolution model capture in practice the growth of urban form, and multiple dynamical regimes between population and road networks. It is calibrated on empirical morphological and network measures computed on spatial windows covering Europe.

We then develop a family of models at the macroscopic scale to simulate systems of cities, and more particularly the co-evolution of cities and interurban transportation networks. Building on dynamical models developed in the frame of Pumain's evolutionary urban theory, this allows investigating self-reinforcement processes between urban and network hierarchies. We show in particular that these models are effectively capturing a diversity of co-evolution regimes (in the sense of a circular causation) between the properties of network and cities.

We finally describe an agent-based model at the metropolitan scale including more complex mechanisms for coupling, in particular a governance process for the growth of the transport network based on game theory, coupled to a Lowry model for land-use dynamics. This model is applied and calibrated on the case study of Pearl River Delta, China.

This synthesis emphasises the role of simulation models in the production of knowledge. Indeed, most results are obtained through the systematic exploration of models and the application of new model validation methods provided by the OpenMOLE platform. Such systematic explorations enable the testing of hypothesis, and clarify theory building when linked to empirical data and stylised facts. This also facilitates a modular approach to modeling, and thus the coupling of concepts and processes from different disciplines.

## A strategy for selecting transit routes to operate in critical situations

Azolin L.[2], Rodrigues da Silva A.[1],
*University of Sao Paulo[1], University of Sao Paulo [2]*

Despite the importance of transport systems for social and economic development of cities, regions and countries, they are susceptible to disruptive events. However, even in the face of some disturbance and subjected to restrictions, transport systems must somehow serve the demand for mobility in the most efficient way.

In this context, this study aims to propose a strategy for selecting priority public transport routes to operate in critical situations. The proposed methodology uses the itineraries of a public transport system and trips from an origin-destination survey data. For each transport route, we identify the traffic analysis zones crossed by the route and the number of trips potentially served along the way. With this approach, a potential transit demand can be obtained. The route with the highest value of potential transit demand is classified as the highest priority route. The calculation of the potential transit demand considers that each subsequent route selected must add the highest potential transit demand to the routes already classified. A source code in Python language was developed to do this iterative analysis.

The method was applied in two urban areas of Brazil, the city of São Carlos and the Metropolitan Region of Maceió. We considered the case of a hypothetical fuel supply shortage in which cars could not be used and the public transport operation would be restricted. Thus, not only trips already done by public transport would be served, but also trips by car that are classified as potential transit demand. Scenarios with different conditions of public transport operation were evaluated. With the case studies, it was possible to identify the potential of the transport system service in each region and to identify zones less accessible than others, for different scenarios. The strategy can be easily applied in different contexts and has great potential to be used as a tool for planning and managing transit systems during critical operational conditions.

## Modelling urban networks using Variational Autoencoders

Kempinska K.[1], Murcio R.[1],
*University College London[1]*

### Introduction:

Cities' transport infrastructure, particularly street networks, provides an invaluable source of information about the urban patterns generated by peoples' movements and interactions. With the increasing availability of street network datasets and the advancements in deep learning methods, we are presented with an unprecedented opportunity to push the frontiers of urban modelling towards more data-driven and accurate models of urban forms.

Temporal and spatial patterns of human interactions shape our cities, making them unique, but, at the same time, create universal processes that make urban structures comparable to each other.

### Variational autoencoder:

Variational Autoencoders (VAEs) have emerged as one of the most popular deep learning techniques for unsupervised learning of complicated data distributions. VAEs are built on top of standard function approximators (neural networks) efficiently trained with stochastic gradient descent (Kingma and Welling, 2014). In this work, we apply VAEs to street network images to learn low-dimensional representations of street networks. We use the representations to make quantitative comparisons between urban forms without making any prior assumptions and to generate new realistic urban forms. A variational autoencoder consists of an encoder, a decoder, and a loss function. In our work, we selected Convolutional Neural Networks (CNNs) (Fukushima 1980; LeCun et al. 1990) as the encoder and decoder architectures.

### Street Networks:

The street networks used for model training and testing were obtained from OpenStreetMap (Haklay and Weber 2008) by ranking world cities by 2015 population from the Global Human Settlement database (https://ghsl.jrc.ec.europa.eu/datasets.php (accessed March 2019)). We saved the street networks as images. As the Variational autoencoders required images to have a fixed spatial scale, we extracted a 3×3 km sample from each city image's center. We resized it to a 64×64 pixels binary image. The final dataset contained 12,479 binary images of 64×64 pixels, which we split into 80% training and 20% testing datasets. During model training, we augmented the training dataset by randomly cropping and flipping the images horizontally. Figure 1 shows images for randomly selected cities.

**Results:**

The trained variational autoencoder minimise the loss function defined in Equation 1. The training is equivalent to minimising the image reconstruction loss, subject to a regulariser. We can inspect the training quality by visually comparing reconstructed images to their original counterparts.

The trained autoencoder learnt mapping from the space of street network images (64×64 or 4,096 dimensions) to a lower-dimensional latent space (32 dimensions). The latent representation stores all the information required to reconstruct the street network's original image, so it is effectively a condensed representation of the street network that preserves all its connectivity and spatial information. These "urban network vectors" can be used to measure the similarity between different street network forms and to perform further similarity analysis, such as clustering. First, we measured the similarity between pairs of vectors as the Euclidean distance. Secondly, we used the urban network vectors to detect clusters of similar urban street forms. We used the K-means clustering algorithm. We identified $K$=3 as the optimal number of clusters for the street image data using the elbow method. The obtained clusters seem to separate street networks based on their street density.

**Conclusions:**

This study is an early exploration of how modern generative machine learning models such as variational autoencoders could augment our ability to model urban forms. With the ability to extract key urban features from high-dimensional urban imagery, variational autoencoders open new avenues to integrating high-dimensional data streams in urban modelling. The study considered images of street networks but could be equally applied to other image data, such as urban satellite imagery. Based on 12,479 city images across the globe, our results showed that VAEs successfully condensed urban images into low-dimensional urban network vectors. This enabled quantitative similarity analysis between urban forms, such as clustering. What is more, VAEs managed to generate new urban forms with complexity matching the observed data. Unfortunately, the resolution of the generated images was low, which was accredited to the small size of the dataset. Future work will repeat model training on a much larger corpus of images to improve the generative quality. Moreover, further work will fine-tune the generative quality by investigating the impact of the size of the latent space (currently fixed to 32 dimensions) and the training objective used (e.g. Wasserstein distance instead of KL divergence).

### Airbnb presence in European cities: a comparative study between Rome, Barcelona and Berlin

Lelo K.[2], Brollo B.[1],
*University La Sapienza[1], University of Roma Tre[2]*

In recent years, the sharing economy has undergone exponential growth among various business sectors. In the tourism and hospitality industry, Airbnb is a notable example of peer-to-peer accommodations, allowing people to rent their unused residential spaces to unknown travellers. Airbnb platform creates a virtual community that benefits both hosts and travellers: hosts earn money by offering rooms, apartments or entire houses through Airbnb on-line platform; travellers gain unique accommodation experiences at affordable prices. In 2019 Airbnb counted more than 7 million listings around the world and its business is increasing through the promotion of other services like guided tours, cooking classes and more. While customers usually benefit from the experience of using Airbnb, there are different questionable aspects that have progressively impacted its popularity, such as the loss of residential population in the most representative neighbourhoods of historical cities, associated with deep socio-economic transformations, excessive touristic pressure and safety issues. Deep transformations are taking place as well in less central residential areas, that are experiencing in recent years the pressure of this economic sector.

Geographic location plays a key role in Airbnb's offer. When choosing amongst the accommodation options posted on the Airbnb platform, customers evaluate both position and price. The latest is strongly influenced by the geographic context: centrality, urban quality and amenities, vicinity to landmarks, accessibility to the transportation system and so on. This influence determines the presence of Airbnb's accommodations spatial clustering in specific parts of the cities.

In this working paper we compare the characteristics and the spatial patterns of Airbnb supply in three European cities analysing the determinants that account for the observed distributions mainly considering census data at the neighbourhood level, to make possible a detailed comparison. Our study areas are Rome, Barcelona and Berlin, three European cities whose metropolitan areas have significant impacts on commerce, culture, politics, tourism, thus attracting millions of travellers for business, leisure, or a combination of the two. As a result, accommodation demand in these cities is high, which makes them preferred places to start Airbnb activities.

We first look at the presence of spatial clusters of Airbnb accommodations and the characteristics of these clusters in the three study areas. After we estimate a spatial regression model using as dependent variable the number of Airbnb's per spatial unit. Empirical results show that the estimated coefficient of the spatially lagged dependent variable is significantly positive in all cases, indicating that the number of Airbnb accommodations in one location is influenced by the number of Airbnb accommodations in neighbouring locations. This enables us to explore the conditions of the different urban *milieus* accounting for these concentrations. We use census data and other integrated data sources to analyse the socio-economic characteristics of the three studied urban contexts. Finally, we compare the policies of the three cities and their vision about future development and potential threats related to mass and fast tourism.

We will present some elaborations about current situation of Airbnb use given the pandemic spread of COVID-19, especially looking at differences in spatial distribution of bookings in the platforms for all the 3 cities considered.

## Inferring the construction year of buildings based on topographic maps

Hamel N.[1], Reichel H.[2],
*University of Geneva[1], swisstopo[2]*

The Swiss Federal *Register of Buildings and Dwellings* (RBD) is a dataset that contains different information about all the buildings in Switzerland. Therefore, it is widely used for analysing urban development and planning. In this context, an important characteristic about a building is its year of construction, which is substantially incomplete in the registry. Moreover, making use of fieldwork to fill this information can be expensive and time demanding.

Unlike the availability of data concerning the year of construction of buildings, topographic maps are common, easy to access, and they present an interesting space-time cover in Switzerland. Moreover, these maps are regularly updated by the *Federal Office of Topography* swisstopo. This way, inferring the year of construction could be based on the fact that if a building appears on a map but not in the previous version, the construction date should lie between the dates of these two maps. Of course, exceptions are possible, as demolitions and reforms can take place, but going through this process manually would be entirely effortful.

In this context, digital image processing can be suitable for two main processes: automatically segmentation of buildings from other features in topographic maps and automatic detection of those buildings along the time gaps. For both cases, image segmentation and object detection can be based on the colour (RGB) bands of the images and other morphological features, allowing houses to be differed from other features. Furthermore, whenever the temporal range covered by the maps could narrow predictions, spatial statistics approaches can be considered, once they contemplate the different spatial patterns present in cities.

Thus, one urges to test whether image segmentation, object detection and spatial statistics techniques can be applied to topographic maps from different years in order to fill gaps in public registry data regarding the year of construction of buildings.

At first, 6km$^2$ squares were used to extract topographic maps and to query for public registry data. Those four areas are from four different cities in Switzerland, which present different urban density and historical formation: Basel (BA), Bern (BE), Biasca and Caslano (TI). In order to avoid making use of different methodologies, only maps presenting enough homogeneity in terms of symbology were considered. A first extraction process obtained binary representations focusing on building footprints out of the large amount of data available on a map. On these *cleaned* maps, both buildings presence detection and buildings morphologic variation detection were performed. Based on these two criteria, a construction year range was determined for each building.

Although, the selected maps were only available until the 1950s. For buildings older than the oldest available maps, a secondary spatial statistics approach was developed to extend the predictions from the first approach. It was based on the hypothesis that the age of the buildings in this period presents spatial dependence and that the urban growth until the 1950s was mainly radial.

The main challenge in this project was finding a way to properly validate the results. *Ergo* the right metric would correctly drive the detection processes. As the construction year in the RBD is the result of a complicated administrative process (and thus often unreliable), it cannot be assumed as a ground truth. Concomitantly, the topographic maps represent other features than the buildings only. Thus, as the RBD and the maps were developed through different methodologies for different goals, finding buildings on which they both agree in terms of construction year is a reliable way of establishing a validation set. This process had to be done manually in order to ensure the reliability of the obtained metric. The described metric was used to determine the rate of success and to characterise the error resulting from the deduction process.

Considering the assignment of the building construction date to a gap between two maps a success rate of 83.9% was obtained. In terms of distance, 84.7% of the building construction years were correctly deduced within a ±5.8 years error margin, according to the known references. To what concerns the spatial interpolator developed for buildings with a construction date older than the oldest available map (1950s), a validation dataset was similarly built. Although, as the national maps were not available in this period, aerial images were used and manually inspected as before. The Root Mean Square Error (RMSE) was computed based on the difference between the year in the RBD database and the interpolated one. By extrapolating this measure to the whole Switzerland, an accuracy of 15.6 years was obtained.

## Using spatial regression approaches and tree-based models to analyse drivers of urban sprawl

Behnisch M.[1], Krüger T.[1], Poglitsch H.[1],
*Leibniz Institute of Ecological Urban and Regional Development[1]*

Dispersed low-density development – "urban sprawl" –leads to more rapid land uptake than compact high-density growth patterns and has many detrimental environmental, economic, and social consequences.

We present and discuss urban sprawl on the planet at multiple scales over a period of 24 years (1990-2014). Urban sprawl has increased most rapidly in highly developed countries. The most sprawled continent is Europe, followed by North America. The spatial pattern of urban sprawl is further described by the Moran's Index (Moran's I) - both global and local. The global Moran's I is the value that expresses whether the entire spatial pattern is clustered, distributed or random. The local Moran's I explores the location of high and low value clusters.

We undertake a comprehensive data inspection and apply several spatial regression techniques as well as tree-based models in order to investigate the interdependency between a dependent variable (e.g., urban sprawl) and several independent variables (e.g., drivers, influential factors). For this purpose, we selected a bundle of case study areas in Germany. The hypothesis is that we see a difference in the explanatory variables of urban sprawl for urban regions as well as, in contrast, in the more rural surrounding regions. Furthermore, typical explanatory variables of urban sprawl will be identified by comparing different case study regions in Germany.

When compiling the database of potential driving forces (influencing factors) at the municipal level, the aim was to obtain a large number of variables taking into account several dimensions (population, mobility, spatial context, land and property market, economy).

Data exploration consists of data inspection, data transformation and correlation analysis. The aim here is to understand the distribution of each variable and to discover dependencies between several variables. All variables are analyzed using histograms, box plots, QQ plots and density estimators. If necessary, a transformation of the variable was performed to model the distribution more accurately. In the further course, it will be shown to what extent spatial regression approaches and regression tree models can be used to investigate the drivers of urban sprawl. To describe the performance of each approach, we measured the mean absolute error (MAE), the mean square error (RMSE) and the coefficient of determination (R2) using cross-validation.

It is important to note, that none of the models of machine learning has made any effort to take spatial autocorrelation into account. Geographically weighted regression (GWR) and Multiscale GWR (MGWR) can be used to explore regional variations in the relationships across the study area (nonstationarity), because it generates local regression models and allows the model's coefficients to vary across the study area. However, it is rather difficult to create a model encompassing a large number of variables for the study areas. In order to devise such a model, the authors stress the importance of employing regression diagnostics, as well as the need to diagnose model collinearity, e.g., variance inflation factor, condition indexes.

The approach is suitable as a basis for further spatial investigations of influential factors at a given point in time as well as for cross-sectional and longitudinal analyses. Knowledge about the relationship between urban sprawl and its driving forces is essential for efficient land use, to guide land-use planning, to define sustainable development strategies and to inform regional planning and management.

## Using machine learning algorithms to detect patterns of urban processes in residential neighborhoods

Sagi A.[1], Broitman D.[1], Gal A.[1], Matyash M.[1], Gez D.[1],
*Technion- Israel Institute of Technology[1]*

Although the physical structure of neighborhoods, as streets or building characteristics, hardly change over the years, socio-economic patterns are dynamic and evolve relatively fast. Some areas experience increasing demand, dwelling or location preferences change, different building types become more attractive, etc. These dynamics create complex and sometimes subtle urban processes spreading over both space and time. However, the traditional urban research usually focuses on one specific process over a short time and small spatial scale. This research suggests a new approach for analyzing urban processes applying big data and machine learning tools. Using more than 23 million housing transaction all over England and Wales during a period of 24 years, we are able to find similarities between urban processes that took place in different locations and times. The operation of finding similarities between elements and group them together is called "clustering". This is an 'unsupervised' learning since no specific process is defined in advance. We find this approach especially suitable for analyzing urban processes since part of them are hidden and difficult to label.

There are common and successful clustering algorithms available. However, in our case we needed to cluster time sequences and no off-the-shelf algorithm can do that. We tested the clustering algorithms K-means and Self Organized Map (SOM) with a 'sliding window' method to simulate a time sequence. We also tested the Hidden Markov Model (HMM) which is a predictive model, but when combined with the Expectation Maximization (EM) algorithm it becomes a clustering algorithm.

The real challenge in the unsupervised learning analysis is the result's validation. In most cases a ground truth for neighborhood processes comparison is not available. Therefore, we cannot set a numerical rate for the accuracy of the models. Data visualizations of time-lapse maps and charts allow to evaluate the models' results according to the relevant literature and prior knowledge. The HMM-EM algorithm showed very stable results and a strong tendency to determine clusters according to geographical characteristics. In contrast, the K-means clustering seems to rely more on the economic and physical characteristics of the neighborhoods. However, the K-means algorithm, although less stable, seems to describe significant urban processes that have taken place across England in the last 24 years. For example, the well-known and extensively researched London gentrification processes. But unlike previous studies, we can also simultaneously compare gentrification processes in other large cities in England. Additional processes that were identified are the suburban dynamics around the large metropolitan areas and the urbanization processes of rural areas in the center of the country.

## Random forest for archaeological predictive modeling: An explorative application to the Canton of Zurich

Castiello M.[1], Tonini M.[2],
*Institute of Archaeological Sciences, University of Bern [1], Institute of Earth Surface Dynamics, Faculty of Geosciences and the Environment, University of Lausanne[2]*

In the present study, we propose an innovative approach seeking to bring together traditional archaeological issues related to the exploration and the analysis of settlement patterns, with the most cutting-edge Machine Learning algorithms and their applications. In more details, starting from a dataset of known archaeological sites dating back to the Roman Age, we applied an ensemble learning method based on decision trees (i.e. Random Forest) to perform Archaeological Predictive Modeling (APM) for the Canton of Zurich, in Switzerland.

APM can be defined as an automated decision-making and probabilistic reasoning tool relevant for archaeological risk assessment and cultural heritage management. Generally speaking, machine learning based approaches are able of learning from data and make predictions starting from the acquired knowledge, through the modeling of the hidden relationships between a set of observations, representing the dependent variable (which are, in our case, the archeological sites), and the independent variables (i.e. the geo-environmental features prone to influence the site locations). The predisposing factors suggesting the presence of roman sites in the area are the environmental features described by: topographic indices derived from the digital elevation model (DEM); different characteristics related to the soil and its aptitude to agricultural activities; strategic and water-related criteria on which past populations may have based their site choice.

The main objective of the present investigation is to assess the spatial probability of presence of Roman settlements in the Canton of Zurich. As results we produced: (1) a probability map, expressing the likelihood of finding a Roman site at different locations; (2) the importance ranking of the geo-environmental features influencing the presence of the archeological sites. These outputs are of paramount importance to verify the reliability of the data and to stimulate experts in different ways. Specialists in both fields (archelogy and data science) are elicited to assess the performances of machine learning based approaches to process archaeological information and to evaluate the benefits and constraints of using such innovative techniques in a non-traditional field of application.

**References:**

Castiello M.E. &Tonini M. (2021) - An Explorative Application of Random Forest Algorithm for Archaeological Predictive Modeling. A Swiss Case Study - 4(1), 110–125. DOI: https://doi.org/10.5334/jcaa.71

## Poverty distribution mapping using a random forest algorithm

Kouwenhoven L.[2], van Aalst M.[3], Koomen E.[2], Dahm R.[1],
*Deltares[1], Vrije Universiteit[2], Vrije Universiteit, Deltares[3]*

Exposure to adverse environmental conditions and natural hazards is often unequally distributed among different socio-economic groups. This seems particularly the case in developing countries. Yet, these countries typically lack the spatially explicit data to study such phenomena and help target relief programmes. To overcome this data gap, we propose a machine learning approach to map the spatial distribution of individuals by socio-economic status.

We utilize a dataset containing detailed socio-economic information for a subset of our region of interest and use it to predict socio-economic status in other regions. We apply a random forest algorithm in combination with several, mainly open data sources such as population density, flood propensity and land use. We expand upon existing methods through the use of crowd-sourced, openly accessible point-of-interest data from Open Street Maps. To analyze the factors that most strongly influence socioeconomic status in a locality, we look at the relative contribution of each individual feature to the trained random forest model, as a proxy of importance.

Our research focuses on two regions that face a similar increase in exposure to natural hazards, but that differ in spatial context and types of available data. The Greater Colombo Area in Sri Lanka is a highly dense urban area, while the larger Mekong River Delta in Vietnam is predominantly rural. In The Greater Colombo Area, we had access to an additional dataset of building footprints. Where such data is available, results are improved. In both cases the expected increase of flooding is likely to strongly affect individuals with a low socioeconomic status. Current methods estimating the impact of these natural hazards treat damage in an absolute sense, potentially undermining the loss faced by the many with few possessions.

Our method is flexible enough to deal with the many regional differences. Additionally, the use of a random forest algorithm allows visualising and analysing the interactions between presence of, and proximity to, different points of interest. Reliable ground-truth data is vital to verify predictions. Where such data was available, our method was generally able to predict average poverty-levels in low-scale units with little error, indicating a high potential and warranting further research.

**A bottom-up approach for delineating urban areas: Optimum criteria for building interval and built cluster size minimising the connection cost of built clusters**

Usui H.[1],
*The University of Tokyo[1]*

Managements of urban public infrastructures are important for sustainable urban space and life. As urban population increases, urban areas are also expanded in order to provide urban residents with their residential places. Since the late 20th century, Japan have experienced it and faced urban sprawl, a set of sparsely distributed buildings. In general, expansion of urban areas accompanies large amount length of road networks in order to provide with residential places linear urban public infrastructure such as water supply system, sewage system, electric power supply and so forth; and their accessibility to anywhere along road networks. However, since the beginning of 21st century, Japan have entered the depopulation era and faced urban shrinkage. In 2014, Japanese central government started to promote local governments to make their urban areas spatially compact in order to reduce their management costs of urban public infrastructure.

However, we face the following two problems. First, a general consensus does not exist for the definition of urban areas and how to delineate urban areas. Since the present urban areas are delineated by merging predetermined basic spatial units (e.g. census units) which meet population density criterion, they depend on how to set basic spatial units and the criterion. Hence, urban areas delineated in this way lack their relation to urban form as the composition of road networks and buildings along them. To investigate this relationship from a morphological point of view, urban areas should be delineated based on the locations of buildings. Second, the relationship between urban areas and their management costs has not been sufficiently investigated. Intuitively, the more urban areas are morphologically compact, the more their management costs can be reduced.

To evaluate whether or not the present and future urban areas are spatially and morphologically compact, a consistent method for delineating urban areas needs to be developed not only focusing on the locations of buildings and road networks connecting them but also considering management costs. In this study, road networks are classified into either global or local systems. National and prefectural road networks are classified as global networks. This is because they play an important role in connecting cities and built clusters based on the regional and national scale planning. Their lengths depend on the location of built clusters. On the other hand, road networks other than national and prefectural road networks are classified as local networks; and play an important role in connecting buildings adjacent to one another in built clusters by roads. In this respect, the relationship between the locations of buildings and total amount of local road networks connecting them should be explicitly formulated in a simple way. Hence, total length of local road networks is modelled as that of the nearest neighbour distance from each building. Their total length depends on both the number of buildings and their locations. Thus, in this respect, management costs for global and local road networks can be regarded as fixed and variable ones, respectively.

In the literature, by modelling road networks which are classified into global and local ones, the optimal criterion of the radius is obtained as the solution of the optimisation problem which minimizes the average total management cost of linear urban public infrastructures and applied to delineate built clusters. Then, built clusters whose size is larger than a criterion (e.g. average size) are delineated as urban areas. In this study, a built cluster is defined as a set of circles whose centre is the location of a building and radius is the nearest neighbour distance of each building (also called interval of adjacent buildings) by judging whether or not each circle has overlapped one another with at least one of the other circles. This is called a clump. The size of a built cluster is defined as the number of buildings in its cluster and called built cluster size. The smaller the criterion of radius is, the more built clusters of small size are delineated. On the other hand, the longer the criterion ofw radius is, although the greater the built cluster size is, the more sparsely distributed buildings are included in built clusters. Hence, there is a trade-off relationship between the average total management cost of both local and global road networks in terms of the criterion of radius.

However, the criteria regarding maximum interval of adjacent buildings and minimum built cluster size have yet to be simultaneously optimised in terms of the average total cost for connecting built clusters. Therefore, the objective of this study is to answer the following research question: how to optimise simultaneously the criteria regarding the maximum interval of adjacent buildings and minimum built cluster size which minimise the average total cost for connecting built clusters? By solving this optimisation problem, we can delineate urban areas which minimise the average total cost of connecting built clusters. This optimisation problem can be applied not only for investigating the relationship between the average total management cost of urban public infrastructures and how to make urban areas spatially compact but also for much broader disciplines which deal with how to cluster spatial entities on the two-dimensional plane which can be decomposed into global and local systems.

**Scaling of urban heat island and nitrogen dioxide with urban population: a meta-analysis**

Wei Y.[1], Caruso G.[1], Lemoy R.[2],
*University of Luxembourg[1], University of Rouen[2]*

Due to urban population growth worldwide, thermal anomalies and toxic air pollution are increasing concern for citizens. Despite this increasing challenge and indications that these environmental problems increase with city size, there is still no consolidated understanding of the effect of city size on urban heat island (UHI) and nitrogen dioxide (NO2) pollution. There is a trend that, nowadays cities are viewed as organisms where their magnitude of socio-economic outcomes changes along with their population size, and these changes can be generalized by scaling laws. However, we find most studies dedicated to UHI or NO2 consider only a single city or analyze a few cities within the top ranks of specific world regions or globally. Therefore, it is necessary to derive the scaling of UHI and NO2 with urban population as a way of quantifying the negative externalities of urbanization.

We intend to fill this gap by conducting a qualitative synthesis of the literature and performing a statistical meta-analysis from published work with the aim to uncover scaling laws of UHI and NO2 with the population size of cities. Under the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guideline, we collect and filter about 500 research outcomes on UHI and NO2 from Scopus and Google Scholar. We select 22 articles at the stage of qualitative synthesis. We summarize them based on their measurement types and geographical locations. Then, to perform meta-analysis, we group those 22 articles and conduct tests of ANalysis Of VAriance (ANOVA) to identify which are the potentially significant model specifications (linear, semi or double log relations). We find logarithmic population size is statistically significant to the maximum UHI intensity, logarithmic maximum UHI intensity, and logarithmic annual mean $NO_2$ surface concentration from 9 articles. We plot those 3 specifications and find out as city grows, to what extent UHI effects and NO2 pollution change with population size.

The qualitative synthesis of UHI studies includes 384 nonduplicated cities from Asia, Europe, North America, and Oceania. Among them, 52 cities are measured by car traverse, 12 cities are measured by monitoring stations, and 328 cities are measured by remote sensing. The qualitative synthesis of NO2 studies includes 1653 nonduplicated cities from Asia, Europe, Africa, North America, and South America. Among them, 4 cities are measured by handy sampler, 61 cities are measured by monitoring stations, and 1588 cities are measured by remote sensing. The ANOVA tests identify significant relations of maximum UHI intensity and logarithmic maximum UHI intensity with logarithmic population size among 52 nonduplicated cities in Asia, Europe, North America, and Oceania. These 52 cities are all measured by car traverse. The ANOVA tests also find significant relation of logarithmic annual mean NO2 surface concentration with logarithmic population size among 24 nonduplicated cities in Asia and Europe. Among these 24 nonduplicated cities, 2 cities are measured by handy samplers, and 23 cities are measured by monitoring stations.

The results of meta-analysis show moving from a city with a population of 100 thousand to a city with a population of 1 million, the max UHI intensity increases by 2.66 °C, the annual mean NO2 surface concentration increases by 14.95 g/m3. Moving from a city having a population of 1 million to a city with a population of 10 million, the max UHI intensity increases by 3.87 °C, the annual mean NO2 surface concentration increases by 21.72 g/m3. Thus, larger cities have higher levels of UHI effects and NO2 pollution. We also give the progress of verifying the NO2 scaling using census data and in-situ and RS-measured NO2 data at the level of Urban Atlas 2012.

## Land value dynamics and the spatial evolution of cities following COVID 19 using big data analytics

Buda E.[2], Broitman D.[2], Czamanski D.[1],
*Rupin Academic Center[1], Technion - Israel Institute of Technology[2]*

In this paper we present results of a land-use forecasting model that we calibrated with vast geo-referenced data of a major metropolitan area. Each land parcel includes information concerning regulations indicating permitted land-uses as well as the certain characteristics of existing buildings. Data concerning all real estate transactions include information about the assets and the price of the exchanges. Based on these data we estimated the spatial dynamics of land values in the metropolitan area over time and identified locations experiencing development pressures. This analysis allows us to forecast plausible futures of the urban spatial configuration. Taking the approach one step further, we propose simulations motivated by the natural experiment of Covid-19. Behind the physical-morphological development of urban areas there are two main players that overshadow all the others: Real estate developers and planners. The adversarial interplay between them is the major force that shapes the dynamic spatial structure of metropolitan areas. In this paper we operationalize this concept by means of four variables: (1) the actual physical shape of the built area (existing buildings) is the playground as it looks right now; (2) statutory plans represent the territorial ordering defined by urban planners, (3) historical land values reflect past preferences of real estate developers; (4) actual land values, in contrast, represent current trends that are likely to be influential in the near future.

First, a large scale regional assessment of land value calculated from the difference between the market price of the property and its construction costs is performed. Then we apply a logit regression intended to reveal and to test the associations between the observed land value density in the last period (the dependent variable) and the following independent variables: (1) the land use defined by the spatial plans, (2) the observed built density and (3) the land value density in the previous period. Finally, we use a kernel density smoothing procedure to create a continuous density surface of three of the variables of interest, the land value during both periods and the building volume.

Using the suggested model, we are able to forecast the future shape of the playground, assuming a continuation of the actual trends. Moreover, modifying the assumptions, it is possible to speculate how different urban spatial trajectories will develop. We apply this to the case of a hypothetical change of land use preferences due to long last impacts of the Covid-19 pandemics. The resulting simulations provide forecasts of the future spatial structure of the metropolitan area. Comparing the actual and the forecasted scenarios we interpret the spatial dynamics of the city as they would be if a business-as-usual-pre-Covid-19 scenario is realized, and possible trend changes if the impact of the pandemic is long lasting.

## On the Impact of deviations from normality on the performance of spatial models

Nilima N.1, Puranik A.2,
*All India Institute of Medical Sciences[1], Manipal Academy of Higher Education[2]*

### Background:

Spatial analysis has been vital in mapping the spread of diseases and assisting in policymaking. Spatial dependence is the spatial relationship of variable values or locations with the neighboring values. The dependence might be observed in the values of response variable, predictor variable, or residual terms. Studies mention various spatial regression models suitable in situations, depending upon where the spatial dependency is observed. The main aim of this study is to investigate the impact that deviations from normality have on the model performance. Effect of queens and rook contiguity weight matrix will also be established.

### Methods:

The Spatial lag, spatial error, spatial Durbin, and spatial Durbin error are the various models considered in the proposed study. A data was simulated for various scenario on the deviations from normality. Several scenarios on extent of deviations from normality were considered and the model performances were studied under each scenario. Model adequacy check was performed using the Lagrange Multiplier (LM) lag, LM error and Robust LM lag, Robust LM error tests. Akaike Information Criterion was used to support the adequacy check evidence. The simulated data findings were then compared to the real data derived from various sources including District Level Household and facility Survey-3 and National Family Health Survey-4 conducted in India.

### Results:

Upon comparison of specified spatial models, *viz.* spatial lag, spatial error, spatial Durbin, and spatial Durbin error models, spatial lag model was found to be most robust to the deviation from normality in most of the scenarios, followed by spatial Durbin error model.

### Conclusion:

In situations where a deviation from normality is observed in the data collected, a spatial lag model can be used to obtain a reliable estimate in order to achieve reliable inference on the desired health related outcome.

## Disease mapping in stray animals using hidden Markov models and spatial modeling

Mayer K.[1],
*ESET[1]*

Early disease diagnosis and subsequent isolation of potentially infected individuals is crucial in contagious disease prevention. Disease progression modeling can not only help understand the evolution of diseases over time but also identify various disease spreading trajectories. This is especially valuable when we need to differentiate between life-threatening and highly contagious diseases that share symptoms—we need to be able to differentiate between them as early as possible.

This is particularly important in overcrowded animal shelters where large populations of young unvaccinated animals are prevalent. In this paper, I use hidden Markov models (HMM) with multidimensional state structures combined with spatial modeling techniques to study the progression trajectories of major contagious diseases in dogs: canine parvovirus, canine distemper, influenza, and corona virus. Because these diseases share symptoms with each other as well as other conditions, they can be challenging to identify and diagnose early. This makes it difficult to isolate infected animals before they spread the virus and causes outbreaks in animal shelters and high mortality rates. Another complication is that some animals brought to shelter are already infected; therefore, we cannot observe all the stages of progression of their disease. Multidimensional state structures of HMM models allow us to investigate dynamics of disease progression and interaction between states over time. This in turn improves model's mapping of the disease progression to the clinical stages of each disease. Multidimensional structure of models allows me to model diseases as the co-evolution of multiple factors—not only progression of symptoms present in animals—but also incorporate spatial information available for each animal. I can account for other factors affecting the animal's health such as location where it was collected, presence of infected animals collected around that time in nearby areas, presence of infected animals in shelter during incubation period, and days in shelter without any symptoms. This allows me to better map disease progression stages to HMM states, and to provide probability estimates of the likelihood well before the animal tests positive for a given disease. Furthermore, I'm able to investigate also the spatial aspect of disease spreading—differentiate where animals get infected most often (in the shelter or outside before collected by the animal services) and identify hotspots of contagion in the city that would benefit most from targeted vaccination programs.

This paper analyses geocoded data collected by the city pound in the city of Daejeon to track the hotspots for contagious animal diseases in the 5th largest metropolitan area in South Korea. Because the symptoms of each collected animal are closely monitored and recorded by the shelter staff, it is possible to identify probabilistically where each animal was infected. As a result, I'm able to identify each animal's infection time and link it to not only a city division but also to an approximate city block. Such identification can help in prevention of outbreaks of these diseases in future—both through recognizing increased likelihood of contagious animals collected in incriminated areas (and their subsequent isolation after arrival to city pound) and by targeted vaccination programs.

## A pattern-based, stochastic approach to analysis and visualization of spatial distribution of race in urban settings

Dmowska A.[1], Stepinski T.[2], Nowosad J.[1],
*Adam Mickiewicz University[1], University of Cincinnati[2]*

Populations in cities in North America, and increasingly in Europe, are mixtures of several distinct groups (frequently racial or ethnic groups). Spatial distributions of various groups differ, leading to a spatially variable mix of the population; there are locations where a single group dominates, and there are locations where they mix to a smaller or larger extent. This results in a complex spatio-racial pattern that needs to be visualized and quantified in order to compare different cities or for a longitudinal study of a single city. Formally, the spatio-racial pattern is a point pattern formed by locations of inhabitants' residences and marked by their races. However, available data consists of a set of aggregated spatial units (census tracts or blocks in the US) that are inhomogeneous. The challenge is to unmix the data and to find a lucid way to visualize and quantify the resultant pattern. We introduce the set of techniques, collectively referred to as the *"racial landscape"* (RL) method, to do just that.

The RL starts by transforming data to a high-resolution (30m) raster grid with each cell containing only inhabitants of a single race. Population in each small aggregation unit (a block in the US), whose spatial distribution within a unit is unknown, is assigned to the unit's constituent cells in such a way that each cell has a homogeneous population. Because the assignment is stochastic, obtained by the Monte Carlo simulation, the pattern within each small unit is arbitrary, but the pattern on larger scales, including the entire city's scale, is not. The result is a grid-based pattern, with each cell having a categorical label (race) and the value of local population density. Such a pattern is straightforward to visualize by assigning different cell colors to each race category and changing the intensity of the color depending on the cell's population density.

The RL quantifies the pattern using a modification of the co-occurrence matrix (CM). CM has a size of $N \times N$ when $N$ is the number of groups. Each entry in the CM is a number of raster cell pairs with a focus cell having label $I$ and the adjacent cell having label $J$. A normalized CM is a joint probability distribution of labels in adjacent pairs of cells. CM is a compact quantification of the racial pattern if population density is not taken into consideration. To account for population density, we introduce the *"exposure matrix"* (EM). EM is calculated in the same way as CM but each adjacency contributes a location-specific value to the matrix instead of the constant value of 1. The contributed value is an average of population densities in the two adjacent cells.

EM is a compact quantification of the racial pattern that takes into account both race and population density. However, many racial demography practitioners need just two numbers to describe the racial character of a city, the diversity metric and the segregation metric. We derive these two metrics from Shannon's Information Theory applied to a bivariate distribution given by the normalized EM. Diversity, a metric that describes a variety of different races (how colorful is the visualized pattern), is quantified by the marginal entropy *H(x)* of EM (entropy of the entire city composition). Segregation, a metric that describes a tendency of same-race cells to aggregate (the pattern has large same color clumps), is quantified by the mutual information *I(y,x)* of EM.

The example of using the RL method is shown in raceland package documentation (https://nowosad.github.io/raceland/articles/raceland-intro3.html). The results obtained by the RL method are illustrated using 2010 US Census data for the Cook County that includes the city of Chicago. The results include the racial landscape visualization and the racial diversity and segregation maps at the local scale (1.8km). The racial landscape shows the distribution of 5 racial/ethnicity groups. The diversity and segregation of the entire Cook County are 1.8393(6) and 0.4732(9), respectively. These numbers are ensemble averages from 100 realizations of Monte Carlo simulations. Numbers in brackets are uncertainties due to the stochastic character of the method; they are very small, affecting only the fourth decimal place of the value.

We suggest that the RL method yields results superior to those which are currently used by racial demographers. It yields an easy-to-understand racial map which quickly conveys the racial character of a city. It also produces maps of local values of diversity and segregation on any desirable scale. In Chicago, we can observe that diversity changes on all scales, but segregation is strongly scale-dependent and manifest itself mostly at large scales. Finally, RL is the only method that can assess the level of segregation without using arbitrary subdivisions of the city.

A computational framework for RL is implemented in the *raceland* R package (https://cran.r-project.org/web/packages/raceland/index.html).

## Super-resolution of satellite imagery using GANs

Schmutz L.[1], Gravey M.[1],
*University of Lausanne[1]*

Satellite imagery evolved significantly over time, in particular the spatial resolution which easily reach 10 m/pixel today. The high variability of this spatial resolution generate inconstancy in studies, especially for the ones working on larger time scales. Therefore, a statistical super-resolution homogenization would allow to extract more information from these lower-resolution images.

With time many approaches were used to make super-resolution. But deep learning revolutionized strategies over the last years. In particular, approaches using GANs show exceptional potential on very diverse task. Therefore, the goal was to investigate the capabilities and limitations of generative adversarial networks (GANs) based single image super-resolution (SISR) for the downscaling of satellite images.

The ESRGAN architecture that represents state-of-the art approach was selected for this task. It is based on the pioneer SRGAN architecture and improves upon it to further increase perceptual quality while massively reducing computational complexity.

Firstly, we created an appropriate dataset composed Sentinel-2A images taken on every continent to encounter a wide variety of scenes. The dataset contains more than 40k image pairs, consisting of a high-resolution image used as the ground truth, and low-resolution counterpart used as the network input. The dataset conception followed the framework used in the SISR field where the low-resolution image is computed from the high-resolution with a factor of 4 in each direction.

Secondly, multiple models were trained in parallel on the specific dataset for more than one million iterations until convergence was reached. The performance of each successful model was evaluated quantitatively and the best performing one was used to compare the performance of a generically trained network against the performance of a specifically trained one.

The satisfactory results encouraged us to test the performance of the trained network on a non-synthetic dataset. For this experiment Landsat 8 imagery was selected. It has natively a similar resolution and is widely used in the scientific community.

## Deep-time snow cover in the Western Swiss Alps: generating long-term Landsat imagery consistent with climate data

Zakeri F.[1], Mariethoz G.[1],
*University of Lausanne[1]*

Mountain regions play an essential role in exchanging energy and water, providing water supply, and influencing regional climate. In these areas, snow cover is one of the most critical parameters influencing water resources, ecosystems, climate, environment, and so on. Remote sensing provides valuable information for observing land use/land cover changes, including snow cover. In many cases, daily snow cover monitoring is necessary to represent the very rapid dynamics of snow. For example, as some mountain grasslands have a short growing season, it is needed to have high temporal resolution maps. However, satellite images do not provide high temporal and spatial resolution snow cover maps due to restrictions such as cloud cover, shadow, and revisiting time. Furthermore, good quality satellite imagery is only available for the last 10-20 years, making long-term studies challenging.

In this study, snow cover dynamics are investigated using Landsat 5,7, 8 and climate data time-series in the Western Swiss Alps. Our research explores the fusion of satellite imagery and physical properties such as temperature and precipitation to generate relatively high spatial resolution daily snow cover maps. The proposed approach synthesizes a set of probable Landsat images using a spatio-temporal analysis of ERA 5 climate data. ERA5 provides daily atmospheric information such as 2m air temperature, 2m dewpoint temperature, total precipitation, mean sea level pressure, surface pressure, and 10m u/v- wind component. ERA5 uses advanced modeling and data assimilation to combine historical observations, including satellite data, into global estimates. In this study, temperature and total precipitation variables are used. The importance of each predictor is determined with a Bayesian optimization algorithm. To evaluate the generated images, a collection of real Landsat images is omitted from the set of known Landsat data, and the proposed method generates synthesized images. Visual comparison and quantitative assessment of synthesized snow cover history in the Western Swiss Alps show a strong agreement with real Landsat images.

### Adjusting spatial dependence of climate model outputs with Cycle-Consistent Adversarial Networks

Francois B.[1],
*LSCE-IPSL[1]*

Climate model outputs often present statistical biases and are commonly corrected using univariate bias correction methods. Most of the time, those 1d-corrections do not modify the ranks of the time series to be corrected. This implies that biases in the spatial or inter-variable dependences of the simulated variables are not adjusted. Hence, over the last few years, some multivariate bias correction (MBC) methods have been developed to account for inter-variable structures, inter-site ones, or both. As proof-of-concept, we propose to adapt a computer vision technique used for Image-to-Image translation tasks (CycleGAN) for the adjustment of spatial dependence structures of climate model projections. The proposed algorithm, named MBC-CycleGAN, aims to transfer simulated maps (seen as images) with inappropriate spatial dependence structure from climate simulations outputs to more realistic images with spatial properties similar to the observed ones. For evaluation purposes, the method is applied to adjust maps of temperature and precipitation from climate simulations through two cross-validation approaches. The first one is designed to assess two different post-processing schemes (Perfect Prognosis and Model Output Statistics). The second one assesses the influence of non-stationary properties of climate simulations on the performance of the MBC-CycleGAN algorithm to adjust spatial dependences. Results are compared against a popular univariate bias correction method, a "quantile-mapping" method, which ignores inter-site dependencies in the correction procedure, and two state-of-the-art multivariate bias correction algorithms aiming to adjust spatial correlation structure. In comparison with these alternatives, the MBC-CycleGAN algorithm reasonably corrects spatial correlations of climate simulations for both temperature and precipitation, encouraging further research on the improvement of this approach for multivariate bias correction of climate model projections.

### Combining global climate models using local information

Mariethoz G.[2], Thao S.[1], Garvik M.[1], Vrac M.[1],
*Laboratoire des Sciences du Climat et de l'Environnement (LSCE-IPSL), France[1], University of Lausanne[2]*

Global Climate Models are one of the main tools used for climate projections. Since many such models exist, it is common to use Multi-Model Ensembles to reduce biases and assess uncertainties in climate simulations and projections. Several approaches have been proposed to combine individual models and extract a robust signal from an ensemble. Among them, the Multi-Model Mean (MMM) is the most commonly used. Based on the assumption that the models are centered around the truth, it consists in averaging simulations of an ensemble, with the possibility of using uniform weights for all models or to adjust the weight to favor some models.

Here, we present a new alternative to reconstruct multi-decadal means of climate variables from a Multi-Model Ensemble, where the local performance of the models is taken into account. This is in contrast with MMM where the weight given to a model is the same for all locations. Our approach is based on a computer vision method called graph cut and consists in selecting for each grid point the most appropriate model, while at the same time considering the overall spatial consistency of the resulting field. The performance of the graph cut approach is assessed and compared to MMM based on two experiments: a first one where the ERA5 reanalyses are considered as the reference, and a second one involving a perfect model experiment where each model is in turn considered as the reference.

We show that the graph cut approach generally results in lower biases than other model combination approaches that do not take into account the local performance of models. At the same time, it preserves a similar level of spatial continuity as MMM approaches. Those results also hold in the case of a weighted MMM, sometimes even when optimal weights are known. Beyond the interest of the graph cuts method demonstrated here, our results show the general need to consider that biases are location-dependent, and indicate that this spatial component can be used to improve the combination or weighting of models.

**Identifying the runoff limits of the Greenland Ice Sheet using the Landsat satellite archive**

Tedstone A.[1], Machguth H.[1],
*University of Fribourg[1]*

Greenland Ice Sheet melting and runoff has accelerated in recent decades, contributing 6 mm of sea level rise since 1992. However, not all meltwater which is produced runs off. At higher elevations underlain by porous snow and firn, some meltwater can percolate into and refreeze within the pore spaces, so the fate of meltwater in these areas is poorly constrained. Here we identify the annual visible runoff limits of the ice sheet for the first time, by mapping surface hydrological features in >25,000 images from the Landsat satellite archive.

We use supraglacial river networks and their surrounding slush fields as proxies for the location of the visible runoff limit. In the percolation zone, rivers and slush fields are sufficiently large to be detectable in Landsat 30m near-infrared (NIR) imagery which is available over the ice sheet from 1985 onwards. We detect river networks by according to their Gaussian-like cross-sections and longitudinal continuity. First, we apply a median filter to reduce noise and speckling. Next, we undertake spectral discrimination with a high-pass Discrete Fourier Transform (DFT) filter which delineates active rivers as dark, linear features with abrupt bright channel banks that manifest as high-frequency information. We then use a Gabor filter preferentially retain linear features followed by a path opening filter to lengthen river channel continuity and further suppress noise. Finally, we apply a threshold to produce a binary hydrology layer.

Subsequently, we filter candidate drainage features to remove outliers then rotate the hydrology layer in the angle of the ice flow direction. Then, in each row of the rotated layer, the leading edge of the first candidate drainage encountered is picked as the location of the runoff limit, yielding a set of up to 8,000 (x,y,z) coordinates which describe the position and elevation of the runoff limit.

These individual runoff limit picks were combined into a PostGIS database containing >33 million rows, which we then reduced through a number of spatio-temporal approaches to deduce trends in the runoff limit behaviour of the ice sheet. Between 1985-1992 and 2013-2020, the runoff limits along the west and north margins rose by 51-334 meters depending on the region, expanding the runoff area by 29%. The largest area increases occurred along the western margin. During the last 10 years the runoff limit has reached high elevations in most melt seasons, enabling runoff even during years of low melt. In total, the expanded runoff area has contributed 601 Gt of runoff to the oceans since 1985.

## High resolution spatial modeling of central Europe's electricity system in 2035 for near-optimal feasible scenarios

Sasse J.-P.[1], Trutnevyte E.[1],
*University of Geneva[1]*

**Overview:**

For decades, electricity system models have been an important tool to probe various electricity system layouts under assumed future scenarios. Traditional models typically account only for national aggregates at low spatial and low temporal resolution, which makes them unsuitable to assess the technical feasibility of electricity systems with high shares of renewable energy sources such as wind and solar, as these are by nature intermittent and spatially variable. In the face of climate change and due to the changing nature of modern electricity systems, electricity system models are required to resolve time and space at an unprecedented high resolution and are therefore becoming increasingly complex and data-intensive. Another challenge arises from the uncertainty of how Europe's electricity system might develop, as the public acceptance of large infrastructure projects such as wind farms is difficult to predict. Thus, electricity system models are required to account for structural uncertainty of future scenarios.

In this study, we address these challenges by accounting for structural uncertainty in the model, and by modeling at high spatial (650 NUTS-3 regions) and temporal (hourly) resolution. We link two optimization models to investigate trade-offs between electricity system costs and the spatial allocation of electricity generation, storage and transmission capacity, for six countries in Central Europe in 2035: Austria, Denmark, France, Germany, Poland, and Switzerland.

**Methods*:***

In our approach, we soft-link two spatially explicit optimization models called EXPANSE and PyPSA. EXPANSE is a bottom-up, single-year electricity system model with annual resolution and it applies a method called Modeling to Generate Alternatives (MGA) to systematically explore maximally different spatial allocation scenarios of electricity generation units with acceptable additional costs. In this study, we allow up to 10% higher than cost-optimal total system costs. PyPSA is an electricity system model without MGA and it complements EXPANSE with hourly power flow computations of variable electricity generation, storage and transmission. The MGA analysis uses 100 scenarios subject to electricity generation, storage, and transmission constraints, and set to achieve the renewable capacity targets derived from each country's nationally determined contributions (NDCs) for 2030.

**Results:**

We visualize the two most extreme spatial allocation scenarios: the allocation with the least system cost and the highest regional equity (most spatially even) in terms of generated electricity. Both scenarios suggest maximizing nuclear capacity in France and Switzerland (by extending plant lifetimes beyond 2035) and expanding offshore wind capacity. The least system cost scenario suggests keeping currently existing capacities of fossil power plants. In Switzerland, the least system cost scenario suggests installing new large gas power plants in the regions of Zurich, Geneva/Vaud and Graubünden, to reduce system costs and dependence on electricity imports. As a result, greenhouse gas emissions are particularly high. Due to the limited amount of investment in additional storage and renewable capacity, all modelled 650 regions are dependent on net electricity imports between countries. Large transmission capacity additions are unavoidable and are especially required within Switzerland (Laufenburg-Zurich-Geneva) and at the borders between all countries.

On the other hand, the most regionally equitable scenario suggests significantly larger investments in solar PV, onshore wind, and battery storage. This leads to 40% less greenhouse gas emissions at a trade-off of up to 10% higher total system costs. We find that France and Poland have large untapped solar PV and wind potentials. Switzerland has a significant rooftop solar PV potential and policymakers should focus their efforts to promote this technology, if a large share of renewable electricity is desired. Grid-scale batteries are primarily allocated in the main supply and demand centers of Central Europe, i.e. north-west and south-east of France, north and south of Germany, and in the west of Poland.

**Conclusions:**

Our analysis indicates that there are significant trade-offs between the costs and the spatial allocation of electricity system infrastructure for Central Europe in 2035: higher regional equity leads to higher total system costs. Conversely, electricity generation units should be concentrated to as few locations as possible to keep total system costs low. Nonetheless, near-optimal spatial allocations can allow some much-needed leeway for feasible technology options in Central Europe's regions. This freedom of choice could potentially lead to higher public acceptance of new installations and lower greenhouse gas emissions but comes at a significant trade-off of up to 10% higher total system costs.

## On the use of mixed-effects models in spatio-temporal machine learning models

Zurita-Milla R.[1], Fakhrurrozi A.[2], Kounadi O.[1], Dimopoulos G.[1], Hernandez A.[1],
*Faculty ITC, University of Twente[1], Research Center for Geotechnology, Indonesian Institute of Sciences[2]*

### Introduction:

Data-driven solutions are becoming pervasive in the geosciences. However, the majority of these solutions – if not all – do not explicitly consider that spatial data is special. This can lead to apparently accurate models that are sub-optimal. For example, spatial autocorrelation in either the target or the explanatory features of regression problems might result in autocorrelated residuals. Such residuals indicate that the selected modeling framework is not ideal. Various geostatistical solutions exist to deal with autocorrelation. However, these solutions cannot deal "natively" with large and/or heterogeneous datasets. Hence more and scientists that use geodata are resorting to the use of machine learning to solve their classification and regression tasks. This change of paradigm requires careful consideration as well as the development of novel solutions that can cope with the complexity of spatio-temporal datasets.

### Coupling mixed effects and machine learning models:

Linear mixed-effects models are mostly used to deal with clustered data. In our context, this clustering refers to data that has a clear hierarchy (e.g. land cover types at multiple conceptual levels) or to longitudinal data (e.g. time series collected from a fixed set of geographical units). From a mathematical perspective, these models rely on two types of features, the so-called fixed and random effects features, to predict the target variable at hand. This formulation allows coupling mixed effects and machine learning models by replacing the fixed-effect term by a non-linear model such as a random forest. This hybrid model is known as a mixed-effects random forest or MERF.

### Case studies:

The possibilities and limitations of MERF models are evaluated using a couple of case studies. In the first one, we model crime at the postal code level in New York City using complaints to the 311 service of the city. In the second one, we use environmental features to model the risk of getting a tick bite in the Netherlands. This model is based on volunteered reports collected via the tekenradar.nl website. All results are benchmarked against the use of a standard RF model and discussed in the context of the need for having a new generation of machine learning models that are geographical by nature.

## Machine Learning based wildfire susceptibility analysis at regional scale: the experience in Liguria, Italy

Trucchia A.[1], Isnardi S.[2], D'Andrea M.[1], Fiorucci P.[1], Tonini M.[3],
*CIMA Research Foundation[1], Politecnico di Torino[2], Université de Lausanne[3]*

Wildfire is a hazardous and harmful phenomenon impacting people and the environment, especially in populated areas. This phenomenon affects human lives, regional and national economies, environmental health, biodiversity, species composition and ecosystems.Wildfires can be considered one of the most complex weather-induced emergencies, exhibiting high non-linearities; it is triggered by several interacting natural and human factors. In the Mediterranean basin, in particular, wildfires are responsible for well documented damaging effects.

During the last years, many efforts on forestry and ecological research have been made with the goal of forest management for a wildfire resilient landscape. Wildfire susceptibility maps are a solid ally in landscape management, since they display the spatial probability of an area to burn in the future, based only on the local proprieties of the given site. They are the first, mandatory step in order to generate wildfire risk maps, once fire severity maps, vulnerabilities and impacts are furnished.

Recent advances in automated learning and simulation methods, like machine learning (ML) algorithms, gave promising results in wildfires susceptibility assessment and mapping. Generally, this quantitative evaluation is carried out by taking into account two factors: the location and spatial extension of past wildfires events and the geo-environmental and anthropogenic predisposing factors that favored their ignition and spreading. When dealing with risk assessment and predictive mapping for natural phenomena, it is crucial to validate the reliability of collected data, as well as the prediction capability of the obtained results. In a previous work (Tonini et al. 2020) authors applied Random Forest (RF) to elaborate wildfire susceptibility mapping for Liguria region (Italy). In the present study, we address to the following outstanding issues, which are still unsolved: (1) the vegetation map used in the previous work included a class labeled "burned area" that masked to true burned vegetation, making the ranking of vegetation types rather difficult; (2) there was the need to compare RF with other ML based approaches; (3) to test the model's predictive capabilities, the last three years of past fires were used, but these are not fully representative of different wildfires regimes, which may span over non-consecutives years.

The results of the proposed work are the following: (1) based on expert knowledge, the class "burned areas" has been reclassified, and the distribution of vegetation types has been updated. This allowed a correct estimation of the relative importance of each vegetation class belonging to this variable; (2) two additional ML based approach, namely Multi-Layer Perceptron (MLP) and Support Vector Machine (SVM), were tested besides RF and the performance and behaviours of each model was assessed, allowing to compare the three ML based approaches and to analyse the pros and cons of each alternative; (3) the training and testing dataset were chosen based on a clustering procedure on the different fire seasons. This accounted for the temporal variability of the burning seasons. As result, our models can perform on average better prediction in different situations, by taking into consideration years characterized by specific wildfire trends. The three ML-based models (RF, SVM and MLP) were validated by means of two metrics: (i) the Area Under the ROC Curve, selecting the validation dataset by using a k-folds cross validation procedure; (ii) the RMS errors, computed via the difference between the predicted probability outputs and the presence/absence of an observed wildfires in the testing database.

**References:**

Tonini, M.; D'Andrea, M.; Biondi, G.; Degli Esposti, S.; Trucchia, A.; Fiorucci, P. A Machine Learning-Based Approach for Wildfire Susceptibility Mapping. The Case Study of the Liguria Region in Italy. Geosciences 2020, 10, 105. https://doi.org/10.3390/geosciences10030105

## Post-fire soil erosion risk map in Portugal: prediction and validation

Parente J.[2], Lopes A.[2], Girona-García A.[2], Basso M.[2], Vieira D.[1],
*Joint Research Centre [1], University of Aveiro[2]*

Wildfires are a recurrent and increasing threat in Mainland Portugal, where over 4,500 thousand hectares of forests and shrublands have burned in the last 38 years. Landscapes affected by those wildfires have suffered an increase of soil erosion processes, which can negatively affect soil carbon storage, reduce fertility, forest productivity, and become a source of pollutants. Taking these in mind, the main objective of this study is to offer a ground base of post-fire soil erosion risk determination for Mainland Portugal, which will provide a set of tools to help forest managers in the post-fire decision-making, and therefore adequately implement mitigation measures to prevent such impacts.

Post-fire soil erosion was assessed by applying the semi-empirical soil erosion model Revised Morgan–Morgan–Finney, to the entire Portuguese forest and shrubland areas according to distinct scenarios (burn severity, climate). This study benefits from the use of several reliable official datasets of soil characteristics, as also from several model calibrations and validation with field data collected in the last 10 years for the 1st and 2nd post-fire years. The obtained soil erosion map identifies areas with higher post-fire erosion risk in the past and for future climate extremes. Findings of this study will be a valuable tool for forest managers to minimize the economic and environmental losses of vegetation fires in Portugal.

**Acknowledgements:**

This work was supported and conducted in the framework of the FEMME project (PCIF/MPG/0019/2017) funded by FCT - Portuguese Foundation for Science and Technology. The study was also supported by: (i) National Funds by FCT - Portuguese Foundation for Science and Technology, under the project UIDB/04033/2020; (ii) National Funds by FCT - Portuguese Foundation for Science and Technology, under the project UID/AMB/50017/2019. Data were provided by the European Forest Fire Information System –EFFIS (http://effis.jrc.ec.europa.eu) of the European Commission Joint Research Centre.

## Semi-Automated Data Processing and Semi-Supervised Machine Learning for the Detection and Classification of Water-Column Fish Schools and Gas Seeps with a Multibeam Echosounder

Minelli A.[2], Tassetti A.[2], Hutton B.[1], Pezzuti C. G.[3], Fabi G.[2],

*Echoview Software Pty Ltd[1], National Research Council, Institute for Marine Biological Resources and Biotechnology (CNR-IRBIM)[2], SFERANET S.R.L.[3]*

Multibeam echosounders (MBESs) are active sonars and have been historically designed for hydrographic purposes, such as submerged obstacle detection, bathymetry and seabed characterization. Recently, water column imaging (WCI) emerged as viable mean for hydrographic data quality control and found increasing use in oceanographic studies. Target detection – where a target is considered an object that demonstrates an acoustic impedance discontinuity across its boundary – for MBES WCI data has, to date, heavily relied on nearly fully supervised and thus time-consuming data processing procedures. In addition, WCI data processing strongly depends on the subjective expertise of the operator that has to select thresholds, extract and then classify targets. Assuming that targets such as fish schools, gas seeps and noise differ in morphological metrics, scattering degree and behaviour in time, the proposed workflow makes use of Echoview Software and ML techniques to: (i) speed-up the target extraction procedure by drafting a generalized workflow that can be run automatically in Echoview; (ii) classify extracted targets using a pre-trained stacking ensemble ML framework.

Acquired WCI used for model testing covers a 1.5*1.5 km wide area surrounding a gas platform in the central Adriatic Sea, where the presence of gas seeps is well known. Similarly, schools of fish are expected to be frequently detected as targets given the function of the platform as artificial habitat. Measurements were made with a Kongsberg EM2040CD dual transducers MBES, hull-mounted and the system configured to transmit 600 μs narrowband acoustic pulses centered at ~300 kHz. Three other surveys (2 on Platform A, 1 on Platform B) were then used as "unseen data" for the evaluation of the proposed classification method. The survey design consisted of 10 evenly spaced, parallel transects.

For MBES data processing, Echoview Software (Echoview Pty Ltd, Hobart, Australia) was used to analyze water column reflectivity data, using a generalised data-processing workflow: MBES pings from both heads were merged into a single variable, from which the bottom depth was automatically estimated; MBES pings were converted to a summarised 2D view of time by range using the Maximum Intensity operator, where each sample (at range R) contains the maximum value of all of the corresponding multibeam samples that are also at range R; statistical algorithms were applied to remove stochastic noise and reverberation; an image analysis algorithm was applied to identify and delineate contiguous clusters of above-threshold backscatter that met specified size conditions on a ping-by-ping basis, which generates "regions" that represent a slice or cross-section through a target; a subset of individual slices were manually reviewed to be used in the following machine learning stage of the analysis, and classified as fish school (FISH), gas seep (GAS), platform leg (PLATFORM) and NOISE (e.g., from side lobes of the transducer or boat wakes).

The file with the labelled slices (and their calculated metrics) for the 1st Survey, that served for training of the ML model, was first prepared by: (i) manually retaining only features (out of the 68 total metrics available), whose relevance was evaluated by common knowledge; (ii) normalizing features to a comparable scale (Yeo-Johnson transformation), with the exception of coordinates; (iii) extending labelling of a slice through the slices belonging to the same insonified object. Then, feature engineering was needed to generate new and sensible features (combination of inter-related features, by Pearson test), before performing the feature selection on the whole set of features by using Mutual Information (MI) scores. A last additional feature was identified based on the results of a K-means algorithm that hierarchically clustered data into 4 groups. The labeled data were then passed to help train the ML procedure with a k-nearest neighbor (kNN) semi-supervised learning called pseudo-labelling, that combined both labeled and unlabeled data to train the following Gradient Boosting Classifier (GBC), resorting to a brute-force grid search to get the optimal hyperparameter setting. Before training the ensemble model, labeled data were randomly stratified while splitting up into training and testing sets (70:30, respectively) in order to have approximately the same percentage of samples of each target class as the complete set.

The overall accuracy of the prediction method was about 98%, while the confusion matrix quantified correct and incorrect final classifications for each of the 4 target classes: 99% for FISH, 97% for GAS, 87% for NOISE and 99% for PLATFORM. From the results obtained over the unseen data we obtained consistent classification results for Survey 1, 2 and 3 where the method was used to evaluate temporal/spatial trends over the same sounded area (Platform A). Predictions on Survey 4 (Platform B) were performed to investigate the spatial transferability of trained models. Results seemed to be consistent, correctly identifying the presence of the unseen platform and the absence of gas seeps that were never manually observed in this relatively small area surrounding Platform B. It leads us to believe that the adoption of the classification model at the new test site could be easily improved by including relatively small site-specific training data.

## A Machine Learning algorithm to nowcast lightning occurrence

D'Andrea M.[2], La Fata A.[4], Amato F.[3], Bernardi M.[1], Procopio R.[4], Fiori E.[2],
*CESI – s.p.a.[1], CIMA Research Foundation[2], UNIL - Institute of Earth Surface Dynamics[3], Università degli studi di Genova - DITEN[4]*

Being responsible for many hazardous events such as human injuries, damages to transmission lines and wind turbines, and wildfires ignitions, cloud-to-ground lightning discharges have been widely studied in literature, often focusing on the microphysical processes in combination with complex atmospheric dynamics. Moreover, many studies confirm the strong correlation between lightning and extreme events, such as hail, tornadoes, and heavy rainfalls. A timely forecast of the lightning occurrence may help early warning activities for safety purposes. Nowcasting the lightning phenomenon is still a great challenge and, considering the complicated interaction between in-cloud and many atmospheric processes, a wide range of approaches is available for such purposes.

This study proposes the use of a Machine Learning algorithm, specifically Random Forest, to perform spatially explicit 1-hour ahead nowcasting of cloud-to-ground lightning occurrence over an area of ~ 35 km$^2$. Precisely, 18 geo-environmental features have been selected and arranged in a 19-dimensional data frame covering the Italian peninsula and the surrounding seas (the 19th dimension representing the target, i.e. the presence/absence of strokes) and have been used to forecast the lightning occurrence of August- October 2018. The choice of the time interval is due to the particularly high intense lightning activity over the area of interest, supported also by the significant precipitation levels and from the fact that, at European scale, August 2018 was the fourth warmest after 1880, 2016, 2017, 2015. The features' importance, derived from the best Random Forest model, shows the agreement between the relationships among variables of the data-driven model and the physically-based knowledge of the phenomenon. The encouraging results obtained in terms of forecasting Accuracy, Probability of Detection and False Alarm Rate confirm the ability, after proper improvements, of Random Forest to properly perform nowcasting of lightning occurrence.

### Visualization and analytics for UK Predatory Bird Monitoring Scheme submissions

Tso M.[1],
*UK Centre for Ecology and Hydrology[1]*

The Predatory Bird Monitoring Scheme (PBMS) is a long-term (established in 1990), national monitoring scheme that quantifies the concentrations of contaminants in the livers and eggs of selected species of predatory birds in Britain. The data derived from this scheme provides important evidence of the potential exposure of wildlife and human populations to environmental contamination. Utilizing a citizen science approach, members of the public send dead carcasses of birds of prey from all over the country. A post-mortem examination is then carried out and a wide range of pollutant concentrations are measured, such as heavy metals, rodenticides, and flame retardants.

The wealth of PBMS data provides a fantastic resource to understand chemical exposure. We have designed an R Shiny application to allow scientists and members of the public to visualize all PBMS submissions over the years with an interactive map (powered by leaflet). The user can filter submissions by years and species, view key pollutant measurements by hovering or clicking on each point, locate their own submissions by providing a submission ID, and run an animation of submissions over the years. To help users understand the context of the spatial distribution of data, the user may overlay custom tiles on the map, such as built areas. In terms of data analytics, our focus is to allow users to have high flexibility to group the data. The users can group data by species, year, provide custom groupings from raster layers (e.g. UK economic zones), or use the clustering algorithm in the app and specify variables for clustering (e.g. latitude, longitude, elevation, distance to coast, and any bird attributes). For each group of data, we then plot graphs of submissions and key concentration values, as well as flagging points that exceed the acceptable range of each group on the interactive map. The acceptable range is defined by users (e.g. 95% prediction intervals). This approach can facilitate greater exploration of the context of the data and identify exceedance values that would not otherwise be flagged by a simple data range check. Ongoing work includes providing more options to filter data (e.g. by cause of death), and allowing users to create interactive risk maps of contaminant exposure over the UK by converting point measurements to maps.

## Assessing the power of crowd-sourced data to map the suitability for valorisation of Green Infrastructure with aromatic and medicinal plants

Pfeifer C.[1], Sommer L.[1], Felder T.[1], Oehen B.[1],
*FiBL[1]*

Aromatic and Medicinal Plants (MAPs) form a promising intervention to improve Green Infrastructure and simultaneously generate income in economic peripheries (Russo et al., 2017). MAPs however have specific requirements and cannot grow everywhere to reach required quality for the industry (Lubbe and Verpoorte, 2011). Understanding suitability of MAPs is critical to prioritize investment and policies.

The Global Biodiversity Information Facility (GBIF), a crowd-sourced data platform for plant occurrence data, is a game-changer in the way data driven suitability of plants can be fitted. Yet, the quality of crowd-sourced data is often not well-defined (Beck et al., 2014). The objective of this research is to assess the potential for using GBIF data to fit ecological niche models in order to identify priority areas for MAPs.

Multivariate Environmental Similarity Surfaces (MESS) (Elith et al., 2010) are fitted with occurrence data for seven specific MAPs (*Arnica montana* L., *Calendula officinalis*, *Gentiana lutea*, *Hypericum*, *Melissa officinalis*, *Mentha*, *Thymus vulgari*s) from GBIF. All georeferenced data in the extent of the Alpine arc are used; Swiss data coverage is relatively low, most data comes from neighbouring countries. The ecological model was fitted with open access biophysical GIS data as predictor, namely data from SoilGrids (Hengl et al., 2017), WorldClim (Fick and Hijmans, 2017), Global Agro Ecological Zones (Fisher et al., 2012) and European digital elevation model (Copernicus Land Monitoring Service, 2016). In addition, an expert-based modelling approach that links suitability criteria to biophysical GIS data is developed based on Swiss growing conditions.

Results from both approaches show similar patterns across Switzerland, suggesting that crowd-sourced data lead to satisfactory results despite of the low coverage of the occurrence data. The seven suitability maps are aggregated to identify hotspots for MAPs. These hotspots are verified based on a number of farm visits and expert interviews in Switzerland and Germany. The results suggest that the MESS mapping approach performs better than the expert based one. Both approaches however fail at identifying a known MAP hotspot in Switzerland. The microclimate in which it is located is not captured in the climatic layer used as a predictor variable.

Suitability maps based on GBIF data leads to good prediction even in areas with low data coverage by including data from a bigger extent. Also, predictive power of the maps can be improved by using more accurate predictor data.

Future work will consist of exploring potential biases in GBIF data by fitting MESS on different subsamples of this data as well as test different predictor data. Also, the inclusion of future climatic data as predictor in these models will allow for identification of shifts in MAPs hotspots resulting from climate change and adaptation of Green Infrastructure.

**References:**

Beck, J., Böller, M., Erhardt, A., Schwanghart, W., 2014. Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecol. Inform. 19, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

Copernicus Land Monitoring Service, 2016. EU-DEM v1.1 —Copernicus Land Monitoring Service [WWW Document]. URL https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1.1 (accessed 9.30.19).

Elith, J., Kearney, M., Phillips, S., 2010. The art of modeling range-shifting species. Methods Ecol. Evol. 1, 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x

European Commission (Ed.), 2013. Building a green infrastructure for Europe. Publ. Office of the European Union, Luxembourg.

Fick, R.J., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37, 4302–4315. https://doi.org/10.1002/joc.5086

Fisher, G., Nachtergaele, F., Prieler, S., Teixeira, E., Toth, G., Velthuizen, H., Verelst, L., Wiberg, D., 2012. Global Agro - Ecological Zones (GAEZ v3.0) - Model Documentation. IIASA, FAO, Laxenburg, Austria and Rome, Italy.

Hengl, T., Jesus, J.M. de, Heuvelink, G.B.M., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., Guevara, M.A., Vargas, R., MacMillan, R.A., Batjes, N.H., Leenaars, J.G.B., Ribeiro, E., Wheeler, I., Mantel, S., Kempen, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLOS ONE 12, e0169748. https://doi.org/10.1371/journal.pone.0169748

Lubbe, A., Verpoorte, R., 2011. Cultivation of medicinal and aromatic plants for specialty industrial materials. Ind. Crops Prod. 34, 785–801. https://doi.org/10.1016/j.indcrop.2011.01.019

Russo, A., Escobedo, F.J., Cirella, G.T., Zerbe, S., 2017. Edible green infrastructure: An approach and review of provisioning ecosystem services and disservices in urban environments. Agric. Ecosyst. Environ. 242, 53–66. https://doi.org/10.1016/j.agee.2017.03.026

## Improving epidemic testing and containment strategies using machine learning

Natali L.[2], Helgadottir S.[2], Marago O.[1], Volpe G.[2],
*IPCF CNR Istituto per i Processi Chimico Fisici[1], University of Gothenburg[2]*

We present a novel approach to optimize testing strategies in the event of large-scale epidemic outbreaks. We employ machine learning to enhance the use of the available resources to identify infectious individuals. These results are reported in detail in [1].

The World Health Organization provides some general guidelines [2] for strategies to prevent disease spreading, including travel restrictions, social distancing, and enforced quarantine. The safest approach would be to isolate and quarantine all individuals; however, this cannot be implemented and maintained on a large scale for a prolonged period because of its societal and economic deleterious effects. An alternative approach is to perform extensive testing to identify and isolate infectious individuals rapidly. It improves our knowledge of the spatial distribution of the disease and the exact number of ongoing cases. The main factors determining the velocity of testing are the population size, the number of available samples, and the time required to carry out the test.

Cost-effective containment strategies rely on making the best possible use of the available resources to identify infects. Our work adopts a simplified model for the population, for the features of the disease and the containment policies. We present a machine learning strategy to select which individuals are most beneficial to test.

We simulate an outbreak using the archetypal SIR model –proposed in 1927 by Kermack and McKendrick [3] –in which the population is split into three groups: Susceptibles, Infectious and Recovered/Removed. The groups respectively include individuals that have never been, currently are, and have previously been infectious. We assume that a small part of the infects is identified as confirmed cases. The latter is employed to train a neural network [4] that learns to make predictions about the rest of the population. The key points are that we perform the training simultaneously to the epidemic outbreak, so the neural network does not have information about the disease features, but only about space and time information regarding the infectious individuals, and we include asymptomatic cases.

We aim to present three main points with this work. We show that –employing the predictions from the neural network –allows us to contain the outbreak more effectively than adopting standard approaches. We compare the overall evolution of the epidemic with and without using the neural network while isolating the same number of individuals. Moreover, we point out how the predictions are less effective if adapted to a different epidemic outbreak, suggesting that the neural network adapts autonomously and dynamically to the underlying SIR model, without explicitly knowing the parameters. Finally, we demonstrate how this method applies when the immunization after recovering is only temporary, so there is the possibility of reinfection (SIRS model). In such a scenario, the neural-network-based strategy can help to eradicate the disease and prevent it from becoming endemic in the population. We envision that similar methods can be employed in public health to control epidemic outbreaks and to eradicate endemic diseases.

**References:**

[1] Natali, L., Helgadottir, S., Marago, O. M. &Volpe, G. Improving epidemic testing and containment strategies using machine learning. arXiv:2011.11717 (2020).

[2] Department of Communications, WHO World-wide. Strategic preparedness and response plan. https://www.who.int/publications/i/item/strategic-preparedness-and-response-plan-for-the-new-coronavirus

[3] Kermack, W. O. &McKendrick, A. G. A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. A115, 115700–721 (1927).

[4] Goodfellow, I., Bengio, Y. &Courville, A.Deep learning (MIT press, 2016).

## Disaggregation of areal unit count data

Anderson C.[1], Lee D.[1], Sanittham K.[1],
*University of Glasgow[1]*

Disease mapping is the field of epidemiology which focuses on estimating the spatial pattern of disease risk across a geographical region. Each actual disease case is associated with a specific location in space, but for confidentiality reasons the exact coordinates of this location are not made available to researcher. Instead, the region of interest is subdivided into a set of administrative districts, and the disease risk data consists of aggregated disease counts at this district level. This means that inference is also generally restricted to estimating disease risk at the district level. Such inference can, however, be susceptible to the modifiable areal unit problem (MAUP), whereby the estimated risk surface can be affected by the arbitrary choice of district boundaries. The administrative districts are often constructed based on historical or political boundaries rather than for clinical reasons, and therefore are not necessarily the most relevant or appropriate for the analysis being carried out. If a different spatial partition of the region was selected, then the counts, and therefore the inference, might appear completely different, even though the true underlying disease data are exactly the same.

In this research, we aim to address this problem by producing "disaggregated" disease risk estimates based on overlaying a regular grid (e.g. 1km x 1km) across the region and instead making inference at the grid level. There have been previous attempts to achieve a similar goal via dasymetric modelling and geographically weighted regression, but in this paper we choose an approach which uses multinomial sampling. In our approach, we redistribute the disease cases from the district level across the grid squares, with weights proportional to population of the grid squares and the areas of intersection between the districts and grid squares. We then fit a conditional autoregressive (CAR) model at the grid level to estimate the spatial risk surface at this resolution.

In addition to tackling the MAUP, this approach is also useful for scenarios where one has two sets of disease data which were observed at different spatial resolutions, since it provides a method for disaggregating both to a common spatial resolution. Another important application is spatio-temporal modelling, where it makes it possible to estimate trends over time, even in scenarios where the administrative boundaries change on a regular basis. Our method is illustrated using an application to respiratory hospital admissions in Glasgow, Scotland.