SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of Economics, Management and Statistics (DEMS)

PhD program in Statistics – Cycle XXXV

Methods for Extracting Valuable Information from Spatial Web and Open Reliable Data

Vincenzo Nardelli

Registration number 854324

Supervisor: Prof. Giuseppe Arbia

Coordinator: Prof. Matteo Manera

ACADEMIC YEAR 2022/2023

Contents

Chapter 1

Introduction

In recent years, there has been a remarkable proliferation of big data across various fields of society. A significant portion of this data, which encompasses open administrative records as well as unstructured data sources such as crowdsourcing, web scraping, and social media platforms, has become publicly available. These days, the pivotal role of data is distinctly highlighted within the so-called artificial intelligence (AI) domain. The proliferation of artificial intelligence applications is largely due to corporate's extensive web data collection rather than just improved methods or computational power. Although neural network architectures are public and replicable, and despite significant corporate investments in computational power, the true differentiator lies in the proficiency of data acquisition from the web and its subsequent refinement to establish robust training datasets. Numerous corporate entities maintain a veil of ambiguity concerning the specifics of their datasets, their methodologies for data collection, and strategies to address inherent biases. This lack of transparency has ignited scholarly debates about the ethical considerations and implications of AI's integration into society.

While harnessing vast quantities of data might seem promising, it isn't always adequate for uncovering all answers in scientific research or practical endeavours. Anderson (2008) once remarked, "With enough data, the numbers speak for themselves," but this notion falls short when confronting real-world challenges. Web and open data, due to their inherent imperfections and biases, shouldn't be directly used into statistical models or analytical tools like machine learning or AI, as they can severely distort empirical findings and skew any conclusions. Alternatively, the methodologies developed so far must be expanded to have the capacity to handle the uncertainty and biases inherent in this type of data.

A significant portion of big data is geocoded , providing a rich reservoir of insights crucial for describing, monitoring, and predicting a wide range of geographical phenomena. Often, these types of data are collected for non-statistical reasons and are characterized by a plethora of imperfections that can significantly

compromise the outcomes of any subsequent statistical analysis. This implies the definition of innovative tools in their collection, processing and modelling such as alternative collection and sampling design or techniques for dealing with bias coming from convenience sample, geomasking, outliers so as to transform them into reliable sources made available for public reuse and modelling, to increase knowledge in far-reaching applications and to support individual and political geo-decisions in a wide range of areas. We refer to the end point of this long analytical process as Spatial Web and Open Reliable Data (SWORD).

The SWORD comprehensive analytical procedure can be delineated into three distinct phases: collection, processing, and modelling. The initial phase, (collection), predominantly pertains to the challenges encountered in sampling. This is where the foundation of our data framework is established, and accuracy at this juncture is paramount to ensure the integrity of subsequent analysis. The subsequent phase, (processing), is focused on addressing two critical issues: the detection of outliers which might skew the results and the practice of geomasking to protect the privacy of location-specific data while maintaining its usability for spatial analysis. Finally, the third phase (modelling) encompasses both the assessment of autocorrelation – a critical factor in spatial analytics given the inherent interdependence of spatial data points – and the subsequent spatial modeling. Each of these phases plays a pivotal role in ensuring the reliability and robustness of the SWORD analytical process. The three phases will be treated in a greater detail in the reminder of this document.

Data Collection

Efficient and accurate data collection forms the backbone of any analytical framework, and SWORD ensures that such data is both statistically and geographically relevant . The COVID-19 pandemic was a clear example of its utility (Arbia and Nardelli, 2020). Indeed, all decision-making processes activated during the emergency phase were based on the use of data not collected for statistical purposes, such as administrative data or crowdsourced data. This integration found a particularly suitable application during the COVID-19 pandemic: as the world grappled with the rapid spread of the virus, there was an imperative need for timely, reliable, and spatially-accurate data to inform policy decisions, track viral spread, and allocate resources. The SWORD framework became instrumental in addressing these challenges, highlighting its pivotal role in modern data-driven decision-making processes (Arbia et al., 2022).

In the paper by Alleva et al. (2022), we introduced an innovative sampling design tailored for building a continuous-time surveillance system, pertinent to the exigent informational demands posed by the COVID-19 pandemic using spatial administrative data. The methodology, characterized by its flexibility to adapt to

evolving stages of an epidemic, rapid operationality essential during emergencies, and statistical optimality properties, emphasizes the importance of real-time, reliable, and geographically comprehensive data collection—values central to the SWORD approach. In second paper (Alleva et al., 2023), we further delved into the challenges presented by the pandemic, proposing an enhanced spatial sampling design that harnesses spatial information and aggregate data to optimize the prevalent two-stage sampling design for studying human populations. This integration of spatial considerations not only amplifies the essence of the SWORD methodology but also introduces a critical evaluation on the tradeoff between efficiency and feasibility. While most of the sampling plans proposed in the literature are grounded in optimal theoretical properties, their practical implementation may pose challenges. As a result, there is a necessity to contemplate suboptimal designs that, while approximating well to the ideal standards of optimality, offer a more feasible and readily applicable approach. This delicate balance between achieving maximum efficiency and ensuring feasibility becomes especially crucial during crisis scenarios, where timely and spatially accurate data is of paramount importance.

Processing

SWORD is instrumental in data processing, especially in domains such as spatial outlier detection and geomasking. Fundamentally, both these procedures address the complexities that arise from spatial data, albeit in different ways. For spatial outlier detections, we assume that the geographical reference is accurate, but discrepancies arise due to the observed value, which might be skewed because of measurement errors or other potential pitfalls. In contrast, geomasking presents the opposite challenge: while the data value is assumed to be correct, its geolocation is intentionally or unintentionally misrepresented. This can be attributed to various reasons, such as technical limitations in the GPS system, errors, or privacy concerns that necessitate obfuscation of the precise location. Leveraging the SWORD framework, some of the papers produced during my doctoral period, have been instrumental in different areas, such as the analysis of price data collected through crowdsourcing and the estimation of spatial models with geo-masked data.

In the first of these papers by Arbia et al. (2023b), we offered insights into the complex task of the international institutions of monitoring food market prices with high spatial and temporal resolution (Solano-Hermosilla et al., 2020). For such organizations, the precision and timeliness of data are paramount, influencing a spectrum of decisions from policy-making to market strategies. The study presents methods for validating data from mobile app-based crowdsourcing within spatio-temporal markets (pre-processing), and subsequently reweight-

ing them weekly based on their geolocation (post-processing). The use of such methodologies ensures accurate, actionable, and time-sensitive data. This was further evidenced by their case study in monitoring food prices in Nigeria, which emphasized the augmented accuracy of their reweighted estimates. In a second paper (Arbia et al., 2023a), we focused on the challenges that arise when preserving confidentiality in data, specifically when point-referenced data, crucial for these spatial tools, is intentionally geo-masked. This geo-masking, while imperative for safeguarding sensitive information, can introduce bias and inefficiencies in spatial predictions and in particular in the kriging modelling. The proposed methodology offers a paradigm wherein data processing accounts for these "intentional locational errors", ensuring that the resultant estimations, even when data is geo-masked, remain robust and reliable.

Modelling

Frequently, methodologies employing this kind of data concentrate predominantly on the dataset's treatment, without advancing towards the intricacies of model estimation. Nevertheless, in the SWORD framework, we elucidate examples even within the domain of spatial autocorrelation and spatial econometrics modelling. Spatial autocorrelation, which essentially extract the degree to which one spatial unit is correlated to its neighbors, can encounter issues from unconventional dataset sampling or the presence of outliers. Such challenges not only distort the immediate analysis but also reverberate through subsequent modelling phases. The new methodologies under the SWORD framework should offer a structured methodology to account for these potential discrepancies, ensuring that the integrity of the spatial data is preserved throughout the modelling process. This enhances the veracity and applicability of the resultant spatial models, making them more adept at capturing real-world spatial phenomena.

Significantly, the proliferation of non-traditionally sourced data, predominantly geo-coded, from alternative sources such as crowdsourcing and web scraping, is reshaping the contours of regional and spatial sciences. In this area we made a contribution with the paper Arbia and Nardelli (2023). In this work we insightfully highlight the pitfalls of treating such data as representative. In fact, they are merely "convenience samples" which do not permit robust probabilistic inferences. In this context, the principles of data reliability inherent in the SWORD framework resonate deeply. The proposed technique to rectify these shortcomings in the spatial inferential context, correcting for biased estimation. In their forthcoming work, Arbia and Nardelli (submitted) further unravel the complexities of spatial micro data by demonstrating that weight matrices in a context of spatial autocorrelation estimation are no longer deterministic. They unveil the inefficiencies that might creep into spatial econometrics models. The insights, especially the re-

alization that empirical spatial lags can be biased estimators, underscore the value of these methodologies in ensuring accurate spatial data interpretation. Lastly, in Arbia et al. (submitted) we explore into methods for robustifying traditional spatial correlation measures amplifies SWORD's potential in optimizing spatial econometrics modelling. By devising mechanisms to counteract observations that might disproportionately skew spatial correlations on a map, they emphasize the need for an approach based on robust estimation of the autocorrelation indexes. This framework's ability to account for varied data sources, outliers, and sampling anomalies makes it a natural fit for the complex world of spatial econometrics laid out in this proposal.

Bibliography

Giorgio Alleva, Giuseppe Arbia, Piero Demetrio Falorsi, Vincenzo Nardelli, and Alberto Zuliani. Spatial sampling design to improve the efficiency of the estimation of the critical parameters of the sars-cov-2 epidemic. *Journal of Official Statistics*, 38(2):367–398, 2022.

Giorgio Alleva, Giuseppe Arbia, Piero Demetrio Falorsi, Vincenzo Nardelli, and Alberto Zuliani. Optimal two-stage spatial sampling design for estimating critical parameters of sars-cov-2 epidemic: Efficiency versus feasibility. *Statistical Methods & Applications*, 32:983–999, 2023.

Chris Anderson. The end of theory: The data deluge makes the scientific method obsolete. *Wired Magazine*, 2008.

Giuseppe Arbia and Vincenzo Nardelli. I dati non parlano da soli: l'epoca del coronavirus smaschera l'inganno dell'algoritmo-onnipotente e rivaluta il metodo statistico. *Giust. Insieme*, 923, 2020.

Giuseppe Arbia and Vincenzo Nardelli. Using web-data to estimate spatial regression models. *International Regional Science Review*, 2023.

Giuseppe Arbia and Vincenzo Nardelli. Robust spatial correlation and spatial outlier detection. submitted.

Giuseppe Arbia, Vincenzo Nardelli, and Chiara Ghiringhelli. Estimating uncertainty in epidemic models: An application to covid-19 pandemic in italy. In *The Economics of COVID-19*, volume 296, pages 105–116. Emerald Publishing Limited, 2022.

Giuseppe Arbia, Chiara Ghiringhelli, and Vincenzo Nardelli. Effects of confidentiality-preserving geo-masking on the estimation of semivariogram and of the kriging variance. *Geographical Analysis*, 55(3):466–481, 2023a.

Giuseppe Arbia, Gloria Solano-Hermosilla, Vincenzo Nardelli, Fabio Micale, Giampiero Genovese, Ilaria Lucrezia Amerise, and Julius Adewopo. From mobile

crowdsourcing to crowd-trusted food price in nigeria: statistical pre-processing and post-sampling. *Scientific Data*, 10(1):446, 2023b.

Giuseppe Arbia, Yasumasa Matsuda, Vincenzo Nardelli, and Junyue Wu. Bias in the estimation moran's coefficient when using web-collected data points. submitted.

Gloria Solano-Hermosilla, Julius Adewopo, Helen Peter, Jesús Barreiro-Hurlé, Giuseppe Arbia, Vincenzo Nardelli, Celso Gorrín-González, Fabio Micale, and Tomaso Ceccarelli. *A quality approach to real-time smartphone and citizen-driven food market price data: The case of Food Price Crowdsourcing Africa (FPCA) in Nigeria*. Publications Office of the European Union, 2020.

Chapter 2

Collection

Spatial Sampling Design to Improve the Efficiency of the Estimation of the Critical Parameters of the SARS-CoV-2 Epidemic

sciendo

Journal of Official Statistics, Vol. 38, No. 2, 2022, pp. 367–398, http://dx.doi.org/10.2478/JOS-2022-0019

Spatial Sampling Design to Improve the Efficiency of the Estimation of the Critical Parameters of the SARS-CoV-2 Epidemic

Giorgio Alleva[1], *Giuseppe Arbia*[2], *Piero Demetrio Falorsi*[3], *Vincenzo Nardelli*[4], *and Alberto Zuliani*[1]

Given the urgent informational needs connected with the diffusion of infection with regard to the COVID-19 pandemic, in this article, we propose a sampling design for building a continuous-time surveillance system. Compared with other observational strategies, the proposed method has three important elements of strength and originality: (1) it aims to provide a snapshot of the phenomenon at a single moment in time, and it is designed to be a continuous survey that is repeated in several waves over time, taking different target variables during different stages of the development of the epidemic into account; (2) the statistical optimality properties of the proposed estimators are formally derived and tested with a Monte Carlo experiment; and (3) it is rapidly operational as this property is required by the emergency connected with the diffusion of the virus. The sampling design is thought to be designed with the diffusion of SAR-CoV-2 in Italy during the spring of 2020 in mind. However, it is very general, and we are confident that it can be easily extended to other geographical areas and to possible future epidemic outbreaks. Formal proofs and a Monte Carlo exercise highlight that the estimators are unbiased and have higher efficiency than the simple random sampling scheme.

Key words: Sampling design; SARS-CoV-2 diffusion; Health surveillance system; Unbiasedness; Efficiency.

1. Background and Purpose

The urgent worldwide need for a method of controlling the spread of SARS-CoV-2 requires an accurate evaluation of the sources of data on which the estimation of the epidemic's main parameters can be based. Only in this way will we be able to monitor the evolution of the epidemic over time while simultaneously supporting decision makers in evaluating the effects of the restrictive measures gradually introduced to try and stop the

[1] Università degli Studi di Roma La Sapienza, Memotef, Via del Castro Laurenziano 9, Rome, 00161 Italy. Emails: giorgio.alleva@uniroma1.it and alberto.zuliani40@gmail.com
[2] Università Cattolica del Sacro Cuore statistical sciences, Piazza Francesco Vito, 1, Rome, 00168, Italy. Email: giuseppe.arbia@unicatt.it
[3] Via di Monserrato 111, Roma, 00186, Italy. Email: piero.falorsi@gmail.com
[4] Università degli Studi di Milano-Bicocca Piazza dell'Ateneo Nuovo 1, Milano, 20126, Italy. Email: vincnardelli@gmail.com
Acknowledgments: We are very grateful to Mike Hidiroglou, Pierre Lavallée and Giovanna Ranalli for their challenging discussion, careful reading of our article, and useful suggestions, all of which have helped us improve the quality of our proposal. We also acknowledge the comments and suggestions received from Francisco Lima and Pedro Campos, President and Director of the Methodology Department at the Portuguese National Statistical Office, and from João Lopes from the same Department.

spread and the time required for the reduction and removal of these measures. In general, this approach enables the production of future forecasts of the evolution of the disease, and these forecasts are the essential basis for achieving an effective healthcare response. Indeed, while some degree of uncertainty is inherent in any statistical model, the level of inaccuracy in terms of monitoring the development of the situation can and must be kept under control.

The objective of the proposed method is the definition of an observational protocol for observing the SARS-CoV-2 epidemic over time and providing statistically unbiased and efficient estimates of the sizes of the different attributes of any population identified as a concern with regard to the epidemic. Moreover, we aim to propose a dynamic monitoring tool that can be suitably calibrated both in the growth phase of the infection rate and in the decline phase, with estimates extended to the parameters of the progressive immunization model for the population. All estimates can be produced with associated reliability measures.

However, apart from a few remarkable exceptions, until now, the data that have been collected favour the examination of cases in which the patients display symptoms. This situation is described in statistics as "convenience sampling", and no sound probabilistic inference is possible under such a sampling approach (Hansen et al. 1953). More precisely, in a formal sample design, the choices of observations are suggested by a precise mechanism based on the definition of the inclusion probabilities of each unit (and, hence, by a sound probabilistic inference method); in contrast, with a convenience sampling, no probabilities of inclusion can be calculated, thus giving rise to over- or under-representation of the sample units.

In particular, several studies on COVID-19 diffusion have clearly shown (e.g., Aguilar et al. 2020; Chughtai and Malik 2020; Li et al. 2020; Mizumoto et al. 2020; Yelin et al. 2020) that the available data strongly underestimate the number of infected people that they are unable to capture, for example, asymptomatic cases with an obvious overestimation of the lethality rate, that is, the number of deaths out of the total number of infected people. On the other hand, a broad data collection method using medical swabs that is carried out on a voluntary basis does not constitute a probabilistic sample either. For instance, the practice of systematically collecting observations from people in the vicinity of supermarkets leads to an over inclusion of healthy people in the sample and to a systematic exclusion of those who (either because they are manifesting symptoms or because they feel weak) have chosen to stay confined at home.

However, it is of crucial importance for government and health officials and for the general population to have a clear understanding of the dynamics of an epidemic while it is in progress so that the government can take appropriate measures and guide individual behaviours. In such a situation, it is essential to set up a data collection system that can provide unbiased estimates and statistically valid comparisons over time and across different geographic areas.

For sampling during an epidemic to be empirically relevant, any data collection design must be technically specified, the properties of the associated estimators have to be proved formally, and the design has also to satisfy the following two conditions:

1. It has to be implemented as a surveillance system (or strictly related to an existing one) and repeated in several waves rather than as a one-time survey.
2. It has to be immediately operational considering the practical implications of the collected data.

The latter point is particularly relevant to the idea that the task may prove challenging, especially in a situation where all health operators are employed full time in emergency operations related to the care of the most severely infected people.

Rather surprisingly, the literature on this subject is still extremely poor. Few contributions have suggested the use of crowdsourced data rather than a sampling design along with officially collected data (Leung and Leung 2020; Sun et al. 2020); the risk of erroneous inferences based on these data has been pointed out by Arbia (2020), Di Gennaro Splendore et al. (2020) and Ioannidis (2020). Our aim is to suggest a sampling design whose statistical optimality properties are formally proven, where the design is also operational and can be immediately put into action upon taking the many practical obstacles that may arise in an emergency into account. Although we have the Italian COVID-19 situation in mind, we are rather confident that the suggested protocol could be easily extended to other countries.

The rest of the article is organized as follows. In Section 2, we present a review of the strategies and experiences in progress with regard to data collection until early April 2020. In Section 3, we present the basic sampling framework of our suggested design by distinguishing two subsets of the population to be surveyed, namely, those in which a state of infection has already been verified and those who were in contact with them (group A) and healthy persons (group B). The different roles of the two groups in monitoring infections during different stages of the epidemic are also discussed. In Section 4, we provide a general description of the sampling schemes for the two groups and the various operations to be realized. In Section 5, we focus on the parameters of interest that we aim at measuring with the suggested sampling design based on the two groups, and we discuss how to disentangle possible overlaps between them whose presence may undermine the statistical properties of the estimations. We prove the unbiasedness of the estimates and derive the expressions of the sampling variances. Section 6 is devoted to envisaging an extension of the proposed methodology to subsequent waves of data collection for the purpose of monitoring phenomena at different moments of time and during different stages of the epidemic. Section 7 illustrates the empirical results of a simulation study. Finally, in Section 8, we suggest some practical implications of the study and future research priorities.

The online supplementary material contains a discussion on the efficiency of the proposed strategy.

2. Data Collection During an Epidemic: A Review of Strategies and Experiences Currently in Progress

In the emergency phase connected with the quick and uncontrolled diffusion of COVID-19, governments and institutions in charge are fully aware that knowledge and understanding of the dynamics at work represent the central element for establishing how to intervene and in which geographical areas intervention is most urgent.

In reviewing the approaches followed by various countries until early April 2020, we can identify four strategies and experiences in progress with regard to the estimation of the disease phenomena in the entire population.

1. The first consists of *massive test campaigns* (regardless of the presence of symptoms) carried out without following a formal sampling design; these are essentially aimed at intervening during outbreaks of the epidemic to identify subjects who are infected but with no symptoms or only slight symptoms. This was the strategy of South Korea and Hong Kong, as well as of the United Arab Emirates, Australia, Iceland, and the Veneto Region in Italy. The limitation of this approach of this approach is the impossibility of making statistical inferences for the whole population based on the results.
2. The second possible strategy consists of *diagnostic tests through a probabilistic sample* according to a planned design for the estimation of the phenomena of interest with predetermined precision levels. This approach is aimed at estimating the effective amount of infections, including those in the asymptomatic population. This approach was used in the project performed by the Helmholtz Center for Research on Infections in Germany; this project was based on testing patients' blood for antibodies to the Covid-19 pathogen and involved over 100,000 individuals (Hackenbroch 2020). Similarly, in Romania, a random sample of 10,500 people living in Bucharest has been planned to detect infected persons by following the directions of the Matei Bals Institute of Infectious Diseases in Bucharest (Romania-insider.com 2020). Finally, a random selection of people who do not meet the testing criteria will be observed at two Canberra locations by the Australian Capital Territory (ABC 2020). All these sample surveys are cross-sectional and useful for measuring the infection rate at a precise instant. However, they have distinct characteristics from those of continuous panel-type surveys with rotated samples for monitoring the evolution of the pandemic over time. This latter type of survey constitutes the proposal of this article. UK and Italy conducted sample surveys at a national level to estimate the real prevalence rate of the infection (ONS 2020; Istat 2020). A critical review of the available data on COVID-19 and on the Italian sample survey project is contained in Alleva and Zuliani (2020) and Alleva (2020).
3. The third strategy consists of a *specific massive web survey* collected from individuals and households that decide to participate on a voluntary basis. Some 60,000 Israelis completed the online daily survey developed by the Weizmann Institute. The participants disclosed personal details, such as their age, gender, address, general state of health, isolation status and any symptoms they may have been experiencing (Rossman et al. 2020). We observed examples of the same strategy in Iceland, Estonia and other countries. The results allow us to compare contagion and testing experiences for people and households with different socioeconomic characteristics. For strategy a), the self-selection mechanism in the sampling process makes it impossible to extend the results to the whole population.
4. Another possible strategy is to use *pre-existing sample surveys* and partially modify them to collect information about the epidemic. Creating an EU 'Corona Panel', which is a standardized European sample test to uncover the true spread of the

coronavirus, is indeed the proposal of the Centre for European Policy Studies, as presented by Gros (2020). The proposal refers, in particular, to the use of the EU-wide sample of the panel of households that participate in regular surveys on economic and social conditions, called the '*EU statistics on income and living conditions*' (EU-SILC). More specifically, Dewatripont et al. (2020) suggested implementing two tests using the EU-SILC panel: the first aimed at assessing whether the subject is currently infected, and the second aimed at testing whether the person has become immune due to previous exposure.

Timeliness is crucial. In this respect, the latter strategy seems to guarantee good results for the European Statistical System (ESS). A quick reflection could be made on the feasibility of inserting additional modules in the questionnaire of the quarterly Labour Force Survey (LFS), obviously in accordance with the data protection authorities.

The International Labour Organization (ILO) has reached out to the National Statistical Offices (NSOs) to understand the impacts of COVID-19 on their statistical operations, particularly in the domain of labour statistics (ILO 2020). The ILO recommended that all countries consider what additional information could be useful for capturing the relevant aspects of the epidemic. NSOs should consider whether some existing topics are of low priority; if so, they can thus be temporarily removed from the surveys to create space for new questions.

Many countries are employing combinations of the previously described approaches for collecting data on the epidemic as well as integrating them with administrative data or other official statistical sources. While sample surveys represent a bedrock for making inferences about the whole population, planning and building integrated informative systems for the epidemic is certainly the right way to attain a deeper comprehension of the phenomenon.

3. The Basic Sampling Framework

In what follows, we aim to propose an observational protocol for the estimation of the number of people infected by SARS-CoV-2 (Alleva et al. 2020). Starting from a population where it has been ascertained that individuals are infected (the population contains *verified* cases), the goal is to estimate the portion of the population that is infected but shows no symptoms (the *asymptomatic* cases). For the purpose of the proposed procedure, the individuals are preliminarily classified into two subgroups of interest, which we refer to as *Group A* and *Group B*.

Group A is the subgroup consisting of individuals for which a state of infection has been verified (they could be either hospitalized or in compulsory quarantine) and of all the people who had contact with them in the previous days. Below, we propose to observe the contacts made up to 14 days before the infection has been diagnosed, with this length being the internationally accepted maximum incubation time. However, the unbiasedness of the sampling strategy we propose is still valid (even if less efficient) if the contacts are reconstructed for a shorter time period (e.g., seven days). Therefore, this group contains all individuals who are foreseen to be infected and not just those for whom their infection status has already been ascertained. Therefore, this group represents both the *apparent* and *latent* dimensions of the epidemic.

Group B contains both healthy people for whom the infection is considered *latent* and those whose infections are still in a phase of incubation, where symptoms can manifest at a future moment in time (up to 14 days later).

The rationale for this breakdown of the population is related to the feasibility of the observational scheme that we propose. Indeed, the proportion of infected people in Group A is much larger than that observed in Group B. Moreover, the number of verified infected people is known through the data collected by health public authorities. Thus, focusing resource investments on observing the contacts of this group maximizes the number of infected people observed in the sample. Nevertheless, it is necessary to observe Group B to produce reliable estimates for the whole population, and this is mandatory for correctly estimating the rate of infected people and the rate of lethality.

Estimates relative to the two subgroups may be obtained on the basis of continuous observations over time and by following two distinct methodologies, both of which based on what is known as *indirect sampling* (Lavallée 2007; Kiesl 2016). Indirect sampling is the same technique that is commonly used for the estimation of rare and elusive populations (Sudman et al. 1988; Thompson and Seber 1996).

It is important to emphasize that the distinctive element of our proposal lies in the estimate of the infected population obtained by combining the results obtained through two samples drawn from populations A and B. This estimate can establish different roles in relation to the various developmental phases of the epidemic (in terms of the sample size and/or type of diagnostic assessment to be carried out).

At the beginning of the epidemic, the infection has the characteristics of rapidity, unpredictability in terms of the level of spread, and apparent concentration in certain geographical areas and categories of subjects. The response of the health system and the containment measures to be used are not yet codified, nor is the behaviour of the population that should be considered "responsible". In this phase, an investigation strategy based on indirect sampling appears to be coherent, with the strategy starting from the immediate surroundings of subjects who have confirmed infections. This is the sampling strategy proposed for Group A that, in addition to the estimation of a rare phenomenon in the population, also provides an immediate (and continuous over time) response to the epidemic where it explicitly manifests itself.

On the other hand, to measure the intensity and the evolution of the phenomenon for large territorial domains and in general with regard to relevant characteristics of people (gender, age, educational qualification, professional status and more), a traditional population panel survey with sample rotation can be carried out for Group B. The survey is associated with an indirect sampling mechanism so that it can trace and sample the individuals who came into contact with the infected people found in this second sample.

This panel survey becomes fundamental during phases that follow the peaks of the epidemic to measure not only the reduction in the number of infections (and therefore to test the positive effects of the containment measures) but also the proportion of the population that had contacts with the virus in the past. During the decline phase of the epidemic (which naturally does not preclude the arrival of new infections in specific territories and environments), the role of the sample from population Group B is fundamental and representative of the entire population followed over time. On the other hand, a diagnostic test must also be identified that takes the relative importance of the infected population and

the population susceptible to infection during the various phases of the epidemic into account. From an operational point of view, it seems convenient to rely on nasopharyngeal swabs for sampling the contacts in Group A, regardless of the phase of the epidemic. For the panel survey, a serological examination may be more convenient, particularly during the declining phase in combination with a part of the sample yet to be evaluated (the swab is also administered to this portion). It is important to emphasize that while the swab allows for an estimation of the infected population at a given moment in time, the serological test allows for the estimation of the portion of the population that had contact with the virus without a time reference. On the other hand, both diagnostic tools provide estimates that are affected by errors, and consequently, the estimates must be considered in probabilistic terms. In particular, to ensure the reliability of the results, while the health protocols for the swab require that the test be repeated over time to ascertain the healing of those who contracted the virus, for the serological examination, diagnostic kits that ensure predetermined levels of specificity and sensitivity can be considered. For a discussion on the impact of these errors in epidemic stages characterized by a different base rates of infection, see Fuggetta (2020).

The combination of the two sampling strategies (with different weights for the ascending and descending phases of the epidemic) represents the competitive advantage of our proposal: it is a dynamic monitoring tool designed to be suitably calibrated both during the growth phase of the infection, providing estimates according to different categories of severity, and during the decline phase, with estimates extended to the parameters of the progressive immunization model for the population.

The advantage of our proposal over a strategy based exclusively on indirect sampling or only on the panel sample can be measured in terms of greater efficiency (and therefore more accurate estimates) and lower investigation costs required to achieve the predetermined levels of precision. In the online supplemental material, we see that the strategy's effectiveness is maximum if Groups A and B have the same size with a large intersection between the two groups. A right choice could be to oversample from group A and obtain a small sample from group B.

4. The Sampling Design

4.1. Population of Interest and Its Breakdown Among the Different Groups

In what follows, let U be the population of interest of size N, and let k ($k = 1,\ldots, N$) denote a person belonging to it. Let v_k be a dichotomous variable that assumes a value of 1 if the state of infection is verified and a value of 0 otherwise. Let $U_v = \{k \in U: v_k = 1\}$ be the subpopulation of U, of size N_v, for whom the infection is verified and let $U_c = U \backslash U_v$ be the complementary subset, of size U_c.

Let y_k be the value of variable y, for person k where y is equal to 1 if the person is infected and 0 otherwise. If $v_k = 1$, then obviously $y_k = 1$; however, if $v_k = 0$, then it is possible that either $y_k = 1$ (an infected person for whom the infection has not yet been verified) or $y_k = 0$ (a healthy person).

The target parameter of our survey, Y, is the total number of infected people (verified or not), that is:

$$Y = \sum_{k \in U} y_k. \tag{1}$$

Let $l_{k,j}$ be a generic entry of a link matrix ($k = 1,2,\ldots,N; j = 1,2,\ldots N$) that is equal to 1 if individual k had contacts with individual j in the past 14 days and 0 otherwise, with $l_{k,k} = 1$ by definition. Starting from U_v, it is possible to define the Group A as:

$$U_A = \left\{ j \in U: \sum_{k \in U_v} l_{k,j} \geq 1 \right\}$$

where U_A includes the subset U_v and all the contacts of the members of that subset.

On the other hand, starting from U_C, it is possible to individuate the group B as:

$$U_B = \left\{ j \in U: \sum_{k \in U_C} y_k l_{k,j} \geq 1 \right\}$$

where U_B includes U_C and all the contacts of the infected people in U_C.

The sets U_A and U_B can obviously overlap. Let us define their intersection as the set

$$U_{AB} = U_A \cap U_B = \left\{ j \in U: L_{vj} L_{Cj} \geq 1 \right\},$$

where

$$L_{vj} = \sum_{k \in U_v} l_{k,j} \text{ and } L_{Cj} = \sum_{k \in U_C} y_k l_{k,j}. \tag{2}$$

The above setup is illustrated in Figure 1 below.

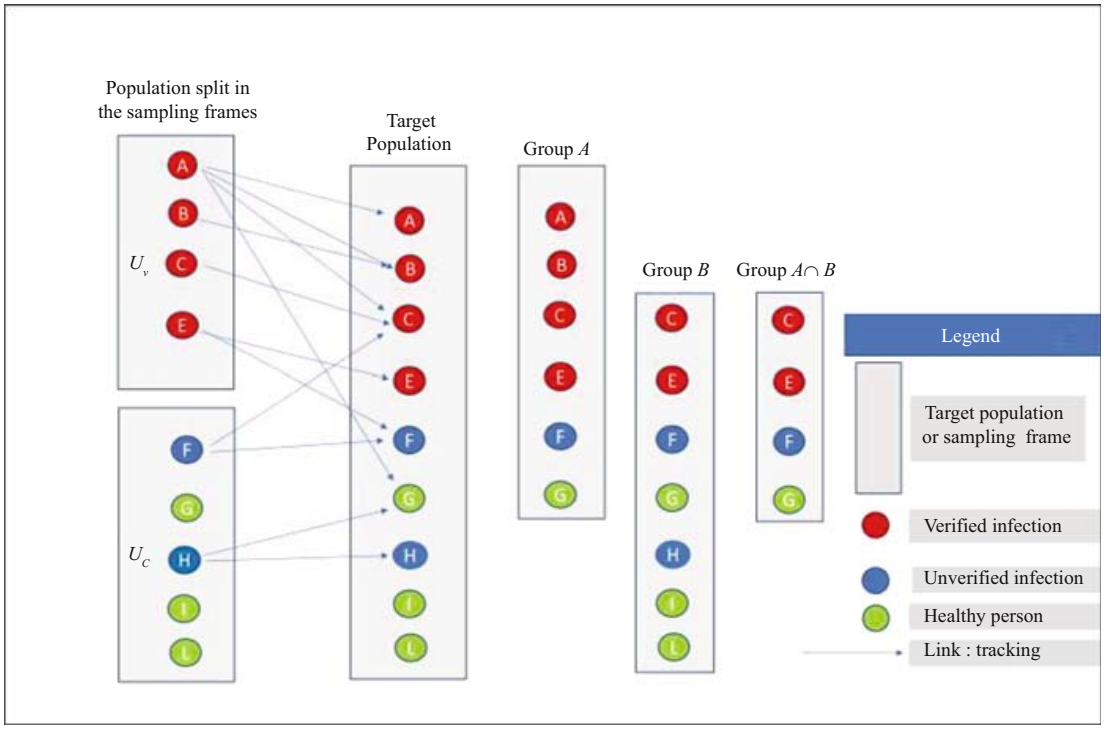

Fig. 1. Population of interest and its breakdown among the different groups.

Two independent samples, namely, S_v and S_C, are selected from the two population subsets U_v and U_C, which represent the sampling frames. The contacts of all infected people in each sample are tracked. The first sample S_v is used to produce an unbiased estimate of the infected people in U_A, while S_C is used to estimate the total of infected people in U_B. The total of infected people in the intersection U_{AB} is estimated from both samples.

4.2. Sampling from U_v

The subset of the people with verified infections increases over time. It is therefore necessary to set up a sampling mechanism that is realized continuously over time. To simplify the sampling description, let us suppose that U_v represents the set of people with verified infections in a given time period. The sampling of U_v is carried out in the following phases:

a) A sample S_v, of size n_v, is selected without replacement from U_v, where the inclusion probabilities are π_{vk} $(k = 1, 2. \ . \ ., N_v)$.
b) All the contacts $U_k = \{j \in U: l_{k,j} = 1\}$ of individual k(selected from S_v) are tracked going back 14 days.
c) A sample S_{vk}, of size n_{vk}, is selected from U_k without replacement and with equal probabilities of inclusion $\pi_{2v|k}$. We use the "2" in $\pi_{2v|k}$ to indicate that this is the inclusion probability of the second stage of the sampling process given the selection of person k in the first stage.

At the end of the above process, the sample $S_A = S_v \cup_{k=1}^{n_v} S_{vk}$ is formed with an indirect sampling mechanism that includes people from both S_v (verified infected people) and $\cup_{k=1}^{n_v} S_{vk}$ (tracked contacts going back 14 days).

The test for verifying an infection is carried out on all the tracked contacts $\cup_{k=1}^{n_v} S_{vk}$. Thus, the value of y, is known for all the people in S_A.

Remark 1. The process of tracking all the contacts of a person can be complex and cumbersome. Different solutions are possible. One possibility is to leverage digital apps, allowing for epidemic control with digital contact tracing, as suggested by Ferretti et al. (2020). Similarly, Ascani (2020) suggested a method based on personal interviews. In this case, the interviewees must be guided in remembering their contacts by means of a specific structure based on the reconstruction of the "social networks" contacted in the days preceding the infection (Scott 2000; Yang et al. 2016).

Remark 2. It is clear that for health and wellbeing reasons and to prevent the spread of the infection, it would be best to examine all infected people. However, from a statistical point of view, obtaining high-quality estimates regarding the number of infected persons is not strictly necessary. From this point of view, it is more important to concentrate effort on repeating the examination regularly over time. The effort required to perform a complete study on the whole population would be unsustainable.

4.2.1. Definition of the Sampling Design

The sampling mechanism for selecting depends on how the data frames for U_v are organized. There are two main possibilities:

Option 1. The data of U_v are available in a centralized data set that can be used for selecting the sample.

Option 2. The data of U_v are available only at a decentralized level so that each healthcare institution has its own list.

The two available options are discussed in turn in the next two subsections.

Sampling Mechanism for Option 1

If the sampling frame of the infected people is centralized in a unified dataset, one could define a *one-stage* design by directly selecting the sample units from the data set. The selection of the sample can be carried out with the cube algorithm (Deville and Tillé 2004, 2005), thus ensuring that the Hortvitz-Tompson estimates (Narain 1951; Horvitz and Thompson 1952) of the selected sample reproduce the known totals of some auxiliary variables (e.g., distribution by sex and age, employment status, geographical distribution, etc.) This can be expressed as follows:

$$\sum_{k \in S_v} \frac{\mathbf{x}_k}{\pi_{vk}} = \sum_{k \in U_v} \mathbf{x}_k, \tag{3}$$

where $\mathbf{x}_k$ is a vector of P auxiliary variables available for unit k

The definition of the optimal inclusion probabilities π_{vk} for indirect sampling that minimize the cost and ensure a predefined level of accuracy for the sampling estimates (or, inversely, minimize the sampling variances for a given budget) can be determined as illustrated by Falorsi et al. (2019). Tillé and Wilhelm (2017) suggested selecting a sample satisfying Equation (3) through a balanced spatial sampling algorithm that is somehow optimal in maximizing the entropy and minimizing the spatial correlations between neighbouring units (Arbia 1994; Arbia and Lafratta 1997, 2002).

Falorsi and Righi (2015) demonstrated that balancing Equation (3) is quite general and allows for the definition of a wide class of sampling designs, including simple random sampling without replacement (SRSWOR), stratified random sampling without replacement (STSRSWOR), stratified random sampling with probability proportional to size (PPS), sampling designs with incomplete stratification (SDIS) and many others.

Assuming that an *SRS* design is used, to obtain the statistical estimates of the number of infected persons in a given *spatial* (the whole national territory or specific geographic area, such as, for example, a region) and *temporal* domain (week/day), it would be sufficient to select approximately 1,000 individuals among the contacts of the infected set of persons for testing. This sample size would ensure a reliable estimate with a coefficient of variation of approximately 5% under the assumption that the proportion of infected people in the target population is approximately 25%.

Sampling Mechanism for Option 2

If the sampling frames for U_v are available only at the healthcare institution level, the selection of units in S_v can be carried out with a two-stage mechanism:

1. First stage. A sample S_{1v} of health care institutions is selected from the population of health care institutions (call it U_{1v}). The first-stage sample is selected without replacement and with PPS, where healthcare institution i is selected with an inclusion probability given by:

$$\pi_{1i} = m\frac{M_i}{M}, \tag{4}$$

in which m is the selected number of healthcare institutions to be included in the first- stage sample, M_i is a measure of the size of unit i and M is the overall measure of size. We may define the measure of size according to different criteria. A good option would be the number of beds available for SARS-CoV-2 patients. The sampling of the health care institutions can be carried out with the already-quoted "cube algorithm", thus ensuring that the Hortvitz-Tompson estimates of the selected first-stage sample reproduce the known characteristics of some auxiliary variables available for the population U_{1v} (e.g., geographical distribution, number of beds available for SARS-CoV-2 patients, etc.). This can be expressed as:

$$\sum_{i\in S_{1v}} \frac{\mathbf{x}_{1iv}}{\pi_{1iv}} = \sum_{i\in U_{1v}} \mathbf{x}_{1iv}, \tag{5}$$

where $\mathbf{x}_{1iv}$ is a vector of auxiliary variables for unit i. As suggested for Option 1, the sample could be selected with a balanced spatial sampling algorithm that is optimal, maximizes the entropy and minimizes the spatial correlations of the neighbouring units. Even in this case, the above balancing Equation allows us to define the general class of sampling designs described by Falorsi and Righi (2015).

2. Second stage. A fixed number of infected people is selected from the sampled institution by *drawing the units* without replacement via a simple random sampling procedure. In such a way, the sampling process is *self-weighting* (Murthy and Sethy 1965) in the sense that all the units in U_v have an equal probability of being selected. Indeed, the final inclusion probability of person k being selected from healthcare institution i is given by the following expression:

$$\pi_{vk} = m\frac{M_i}{M}\frac{\bar{n}}{M_i} = m\frac{\bar{n}}{M}, \tag{6}$$

where $\bar{n}$ denotes the fixed number of infected people selected from each sampled institution.

The *self-weighting* property defines a sampling design that is somehow optimal (Kish 1965) in the sense that it avoids the negative impact of the variability of the sampling weights on the sampling variances.

The sampling selection criterion could be based on a time mechanism, as this is feasible and easily implementable at a decentralized level. For instance, a sample of infected people could be selected by considering those who had access to the healthcare institution within a two-hour time period.

4.3. Sampling from U_C

In this subsection, we illustrate the sampling design for the first selection process, where we sample a panel of individuals independently from S_v for estimating the total of infected people in U_B. Afterwards, we monitor these people repeatedly over time.

The operational aspects to be carried out in this first sampling process are as follows:

a) First, a sample S_c, of size n_c, is selected without replacement from U_c, where the inclusion probabilities are $\pi_{ck}(k = 1,2. \ . \ ., N_c)$.

b) The people in the panel take a diagnostic test on a regular basis (for example, once a month). If member k of the panel receives a positive test result (i.e., $y_k = 1$), all their contacts U_k are tracked up to 14 days back in time.
c) If $y_k = 1$, a sample S_{Ck}, of size n_{Ck}, is selected from U_k without replacement and with equal inclusion probability $\pi_{2C|k}$. We adopt $\pi_{2C|k}$ for the second stage inclusion probability, where the same notation as that of $\pi_{2v|k}$ is used. At the end of the whole process, the sample $S_B = S_C \cup_{k=1:y_k=1}^{n_C} S_{Ck}$ is formed with an indirect sampling mechanism, including people from both S_C (people for whom their infection statuses are not known) and $\cup_{k=1:y_k=1}^{n_C} S_{Ck}$ (tracked contacts of the infected people in S_C, going back 14 days).

Remark 3. The populations U_v and U_C change as a function of time. The panel can be representative of the shifting population. We discuss this topic later on in Section 6. Here, we note that in the subsequent surveys, the verified infected people in the panel are automatically captured by the sampling mechanism defined for the population U_v. However, sample S_C is smaller in size than the total population, observing only the nonverified infected people. This reduction in the sampling size makes it necessary to regularly refresh the panel over time.

4.3.1. A Note on Some Practicalities of the Sampling Design

The sampling design of the panel can be carried out according to different schemas, depending on the availability of the frame and on other organizational aspects. One possibility is to form a subsample from a regular survey of households carried out by official statistics. Here, we assume that the frame of U is represented by a register that is available at a central level and that for each sample unit, we form a set of auxiliary variables. Furthermore, we assume that in this register, the subset U_C can also be identified.

In this informative context, a one-stage sampling design can be carried out with optimal inclusion probabilities π_{ck}, as determined following the steps of Falorsi and Righi (2015), Falorsi et al. (2019). The sampling can then be carried out with a balanced spatial sampling algorithm (Tillé and Wilhelm 2017), thereby ensuring that the following balancing equations are satisfied:

$$\sum_{k \in S_C} \frac{\mathbf{x}_k}{\pi_{ck}} = \sum_{k \in U_C} \mathbf{x}_k. \tag{7}$$

Remark 4. The *panel* can be constructed using a two-phase design so that the selection process can be executed by *pre-screening* two sub-groups, namely:

1. A number of individuals who continue to travel (and are therefore more subject to being infected than individuals who are not traveling).
2. A number of individuals with few contacts who fully observe the prescribed quarantine recommendations.

Remark 5. The two-phase mechanism could be useful if the identification of U_C can not be carried out. This can be realized in the two-step pre-screening phase.

The number of persons involved in the *panel* may be approximately 1,000 (to obtain approximately 1,200 tested individuals) for a given *territorial* and *temporal* sampling domain, thus guaranteeing a reliable estimation with a coefficient of variation of approximately 10% (assuming that the proportion of infected people in this target population is approximately 10%).

4.4. Final Comments on the Sampling Design

We first note that in our proposal, we subsample from the determined list of contacts. We adopt this choice for controlling the costs associated with the survey. However, we could extend the sample to all sets of contacts. Furthermore, if we continue tracking the contacts until all the people being tracked are not infected, the adopted sampling design becomes a classic adaptive schema (Thompson and Seber 1996), which can thus be seen as a particular application of our proposal.

Given the complexity of the epidemiology of COVID-19, it may be useful to consider subgroups in Group B. This may become useful based on the need to consider heterogeneity in the population. In particular, it may be important to consider breaking down certain epidemiological parameters into different subgroups (e.g., transmission coefficient, time to become infectious, proportion of detected cases, time from infection to detection, time to recover). Therefore, we suggest defining four subgroups considering two binary factors: low-risk/high-risk and low-mobility/high-mobility. These subgroups are described as follows:

1. A number of individuals not belonging to high-risk groups who continue to travel/work (and are therefore more subject to being infected and infectious than non-travellers).
2. A number of people not belonging to high-risk groups with few contacts who fully observe the prescribed quarantine recommendations.
3. A number of individuals belonging to high-risk groups who continue to travel/work (and are therefore more subject to being infected and infectious than non-travellers), such as health-care workers.
4. A number of people belonging to high-risk groups with few contacts who fully observe the prescribed quarantine recommendations.

The subgroups may be identified using a two-phase design so that the selection process can be executed by *pre-screening* four sub-groups. For Group A, there might be some advantage in considering the same four subgroups, since the transmission coefficient of each of these subgroups can be significantly different from the others.

Considering four subgroups in both Groups A and B may impact the sample size required to obtain a given sampling error at the subgroup level. Group B has the potential for enabling the study of some crucial "invisible" parameters of the epidemiology of COVID-19 (e.g., the proportion of asymptomatic cases, the time for symptomatic and asymptomatic people to become infectious, and even the proportion of undetected symptomatic cases) in detail. This is also true for each of the four subgroups independently. The sample size for Group B should be defined with this in mind. Population density is also an important factor to control when designing the sampling process.

5. Sample Estimation of the Total Number of Infected People

The total number of infected people Y for each time, and each territorial unit may be expressed according to the breakdown of U, among the different groups (U_A, U_B, and U_{AB}), as

$$Y = Y_A + Y_B - Y_{AB}, \tag{8}$$

where

$$Y_A = \sum_{k \in U_v} \sum_{j \in U} \frac{1}{L_{vj}} l_{k,j} y_j, \tag{9}$$

$$Y_B = \sum_{k \in U_C} y_k \sum_{j \in U} \frac{1}{L_{Cj}} l_{k,j} y_j, \tag{10}$$

$$Y_{AB} = \sum_{j \in U : L_{vj} L_{Cj} \geq 1} y_j, \tag{11}$$

in which L_{vj} is a quantity introduced to control for the *multiplicity* of the measurement of y_j among the different k units of U_v in Equation (9); L_{Cj} is a quantity introduced to control for the *multiplicity* of the measurement of y_j among the different k units of U_C in Equation (10). We may obtain alternative expressions of Y_{AB} starting from the sampling frames of U_v and U_C:

$$Y_{AB} = \sum_{k \in U_v} \sum_{j \in U} \frac{1}{L_{vj}} l_{k,j} y_j \mathbb{I}(L_{Cj} \geq 1), \tag{12a}$$

$$Y_{AB} = \sum_{k \in U_C} y_k \sum_{j \in U} \frac{1}{L_{Cj}} l_{k,j} y_j \mathbb{I}(L_{vj} \geq 1) \tag{12b}$$

where $\mathbb{I}(x)$ equals 1 if x is true and 0 otherwise. The expressions (12a) and (12b) are useful during the estimation phase, as illustrated in Subsection 5.3.

We can compute a direct estimation of the total number of infected people Y for each time and each territorial unit as:

$$\hat{Y} = \hat{Y}_A + \hat{Y}_B - \hat{Y}_{AB}, \tag{13}$$

with

$$\hat{Y}_{AB} = a\hat{Y}^A_{AB} + (1 - \alpha)\hat{Y}^B_{AB}, \tag{14}$$

where $\hat{Y}_A$ and $\hat{Y}^A_{AB}$ are the generalized weight share method (GWSM, Lavallée 2007) estimates of the totals Y_A and Y_{AB} derived from the sample S_A; $\hat{Y}_B$ and $\hat{Y}^B_{AB}$ are the GWSM estimates of the totals Y_B and Y_{AB} calculated from the sample S_B; and $\hat{Y}_{AB}$ is a convex combination of the GWSM estimates $\hat{Y}^A_{AB}$ and $\hat{Y}^B_{AB}$, with $0 \leq \alpha \leq 1$. The parameter a can either be fixed in advance or calculated from the survey data. Further discussion on the choice of α is provided in Subsection 5.3.

5.1. *Estimation of the Component* $\hat{Y}_A$

The GWSM estimator of the total number of infected people in group A, as expressed in Equation (9), is given by:

$$\begin{aligned}\hat{Y}_A &= \sum_{k\in S_v}\frac{1}{\pi_{vk}}\sum_{j\in S_{vk}}\frac{1}{\pi_{2v/k}}\frac{1}{L_{vj}}l_{k,j}y_j \\ &= \sum_{k\in S_v}\frac{1}{\pi_{vk}}\hat{Z}_{vk},\end{aligned} \tag{15}$$

where

$$\hat{Z}_{vk} = \sum_{j\in S_{vk}}\frac{1}{\pi_{2v/k}}\frac{1}{L_{vj}}l_{k,j}y_j \tag{16}$$

represents the second-stage estimate of

$$Z_{vk} = \sum_{j\in U_k}\frac{1}{L_{vj}}l_{k,j}y_j. \tag{17}$$

Remark 6. The term L_{vj} in the previous equation corresponds to the total number of contacts of unit j with people who have verified infections. It can be collected either with digital contact tracing (Ferretti 2020) or by interviews.

Proof of the unbiasedness of $\hat{Y}_A$

This proof can be found in Subsection 5.1 of Lavallée (2007). Denoting the sampling expectation operator as $E(\cdot)$, we have

$$E(\hat{Y}_A) = E\left[\sum_{k\in U_v}\sum_{j\in U}\frac{\delta_{vk}}{\pi_{vk}}\frac{\delta_{2vj|k}}{\pi_{2v|k}}\frac{}{L_{vj}}l_{k,j}y_j\right], \tag{18}$$

where δ_{vk} is a dichotomous variable with $\delta_{vk} = 1$ if $k \in S_v$ and $\delta_{vk} =$ otherwise. $\delta_{2vj|k}$ is a second dichotomous variable with $\delta_{2vj|k} = 1$ if $j \in S_{vk}$ and 0 otherwise.
From Equation (18), we obtain:

$$E(\hat{Y}_A) = \sum_{k\in U_v}\sum_{j\in U}\frac{E(\delta_{vk}\delta_{2vj|k})}{\pi_{vk}\pi_{2v|k}}\frac{1}{L_{vj}}l_{k,j}y_j. \tag{19}$$

However, since:

$$E\left(\delta_{vk}\delta_{2vj|k}\right) = E\left[\delta_{vk}E\left(\delta_{2vj|k}|\delta_{vk}=1\right)\right] = E\left[\delta_{vk}\pi_{2v|k}\right] = \pi_{vk}\pi_{2v|k}, \tag{20}$$

plugging Equation (20) into Equation (19), we finally obtain:

$$E(\hat{Y}_A) = \sum_{k\in U_v}\sum_{j\in U}\frac{1}{L_{vj}}l_{k,j}y_j = Y_A. \qquad \text{Q.E.D.}$$

Variance of $\hat{Y}_A$

The main results on this topic can also be found in Subsection 5.1 of Lavallée (2007). On the basis of the theorem on two-stage sampling (Cochran 1977), the variance of $\hat{Y}_A$ can be expressed as follows:

$$V(\hat{Y}_A) = V_1\left(\sum_{k\in S_v}\frac{1}{\pi_{vk}}Z_{vk}\right) + \sum_{k\in U_v}\frac{1}{\pi_{vk}}V_2\left(\sum_{j\in S_{vk}}\frac{1}{\pi_{2v|k}}\frac{1}{L_{vj}}l_{k,j}y_{j,}\right). \tag{21}$$

In the previous expression, the variance is decomposed into the sum of the first-stage variance (V_1) and the first-stage expectation of the second-stage variance (V_2). All the elements of the previous expression can be estimated with standard statistical inferential techniques (see Horvitz and Thompson 1952; Kish 1965).

5.2. *Estimation of the Component* $\hat{Y}_B$

The GWSM estimator of the component $\hat{Y}_B$ is given by:

$$\begin{aligned}\hat{Y}_B &= \sum_{k\in S_C}\frac{1}{\pi_{Ck}}y_k\sum_{j\in S_{Ck}}\frac{1}{\pi_{2C|k}}\frac{1}{L_{Cj}}l_{k,j}y_j \\ &= \sum_{k\in S_C}\frac{1}{\pi_{Ck}}\hat{Z}_{Ck}\end{aligned} \tag{22}$$

where the term

$$\hat{Z}_{Ck} = y_k\sum_{j\in S_{Ck}}\frac{1}{\pi_{2C|k}}\frac{1}{L_{Cj}}l_{k,j}y_j \tag{23}$$

represents the estimate of

$$Z_{Ck} = y_k\sum_{j\in U_k}\frac{1}{L_{Cj}}l_{k,j}y_j. \tag{24}$$

Proof of the unbiasedness of $\hat{Y}_B$

To prove the unbiasedness of $\hat{Y}_B$, we start with:

$$E(\hat{Y}_B) = \sum_{k\in U_C}y_k\sum_{j\in U_k}\frac{E(\delta_{Ck}\delta_{2Cj|k})}{\pi_{Ck}\pi_{2C|k}}\frac{1}{L_{CJ}}l_{k,j}y_j, \tag{25}$$

where δ_{Ck} is a dichotomous variable with $\delta_{Ck} = 1$ if $k \in S_C$ and $\delta_{Ck} = 0$ otherwise. $\delta_{2Cj|k}$ is a dichotomous variable with $\delta_{2Cj|k} = 1$ if $y_k = 1 \cap j \in S_{Ck}$ and 0 otherwise.
However, we have:

$$E\left(\delta_{Ck}\delta_{2Cj|k}\right) = E\left[\delta_{Ck}E\left(\delta_{2Cj|k}|\delta_{Ck}=1\right)\right] = E\left[\delta_{Ck}\pi_{2C|k}\right] = \pi_{Ck}\pi_{2C|k}. \tag{26}$$

From Equations (25) and (26) it follows that:

$$E(\hat{Y}_B) = \sum_{k\in U_C}y_k\sum_{j\in U}\frac{1}{L_{CJ}}l_{k,j}y_j = Y_B. \qquad \text{Q.E.D.}$$

The term L_{Cj} corresponds to the total number of contacts of unit j with people who have unverified infections. Similar to the estimation process of $\hat{Y}_B$, this information can be collected either with digital contact tracing or by interviews. Alternatively, we can

determine L_{Cj} by following a *back-tracing process*, if unit j is infected, we should test the all their contacts for COVID-19.

Variance of $\hat{Y}_B$

The variance may be obtained by simply adapting expression (21):

$$V(\hat{Y}_B) = V_1\left(\sum_{k\in S_C}\frac{1}{\pi_{Ck}}Z_{Ck}\right) + \sum_{k\in U_C}\frac{1}{\pi_{Ck}}V_2\left(y_k\sum_{j\in S_{Ck}}\frac{1}{\pi_{2C|k}}\frac{1}{L_{Cj}}l_{k,j}y_j\right). \tag{27}$$

5.3. Estimation of the Component $\hat{Y}_{AB}$

Starting from expression (12a), by using the data from sample S_A, we obtain the GWSM unbiased estimator of Y_{AB} as:

$$\hat{Y}^A_{AB} = \sum_{k\in S_v}\frac{1}{\pi_{vk}}\sum_{j\in S_{vk}}\frac{1}{\pi_{2v|k}}\frac{1}{L_{vj}}l_{k,j}y_j\mathbb{I}(L_{Cj}\geq 1). \tag{28}$$

Starting from expression (12b), by using the data from sample S_B, we derive the GWSM unbiased estimator of Y_{AB} as:

$$\hat{Y}^B_{AB} = \sum_{k\in S_C}\frac{1}{\pi_{Ck}}\sum_{j\in S_C}\frac{1}{\pi_{2C|k}}\frac{1}{L_{Cj}}l_{k,j}y_j\mathbb{I}(L_{vj}\geq 1). \tag{29}$$

The information about the intersection of the samples with the subpopulation U_{AB} may be collected either during the interviews or with digital contact tracing.

Singh and Mecatti (2011) provided an in-depth illustration of the different approaches in the literature that are used find the optimal value of α in the context of multiple frame surveys. Hartley (1962, 1974) proposed choosing α in Equation (14) to minimize the variance of $\hat{Y}$. Because the frames are sampled independently, the variance of $\hat{Y}$ is:

$$\begin{aligned} V(\hat{Y}) =& V(\hat{Y}_A) + V(\hat{Y}_B) + \alpha^2 V\left(\hat{Y}^A_{AB}\right) + (1-\alpha)^2 V\left(\hat{Y}^B_{AB}\right) + \\ &- 2\alpha Cov\left(\hat{Y}^A_{AB}, \hat{Y}_A\right) - 2(1-\alpha)Cov\left(\hat{Y}^B_{AB}, \hat{Y}_B\right). \end{aligned} \tag{30}$$

Thus, for general survey designs, the variance-minimizing value of α is:

$$a^{opt} = \frac{V(\hat{Y}_B) + Cov\left(\hat{Y}^B_{AB}, \hat{Y}_B\right) - Cov\left(\hat{Y}^A_{AB}, \hat{Y}_A\right)}{V(\hat{Y}_A) + V(\hat{Y}_B)}. \tag{31}$$

Unfortunately, the above quantity depends on the variable y.

Note that if one of the covariances in Equation (31) is large, it is possible for α^{opt} to be smaller than 0 or greater than 1. Hartley (1974) suggested opting for this alternative expression:

$$a^* = \frac{V(\hat{Y}_B)}{V(\hat{Y}_A) + V(\hat{Y}_B)}. \tag{32}$$

Unbiasedness and variance. The proof of unbiasedness and the calculation of the variance of the estimator $\hat{Y}_{AB}$ are straightforward extensions of what has been illustrated in Subsections 5.1 and 5.2.

Remark 7. Lavallée and Rivest (2012) proposed estimating the total Y with the *generalized capture-recapture estimator* (GCRE), which makes joint use of the capture-recapture *Petersen estimator* and GWSM estimators. In our context, the GCRE estimator may be expressed as:

$$\hat{Y}_{GCRE} = \frac{\hat{Y}_A \hat{Y}_B}{\hat{Y}_{AB(S_A \cap S_B)}}, \tag{33}$$

where $\hat{Y}_{AB(S_A \cap S_B)}$ is the estimate of Y_{AB} that is computed on the basis of the units observed in the intersection $S_A \cap S_B$. The sampling weights for producing the estimates from $S_A \cap S_B$ are given in formula (11) in the abovementioned paper. With respect to expression (33), the GCRE estimator allows for estimating the hidden population that would not be visible with either the public health structure or with the panel survey (e.g., the people who died at home), and this group is very difficult to capture with the usual survey techniques. The main problem for adopting the GCRE estimator is that it would require an overlap of the samples of Groups A and B.

Remark 8. In Section 7 and in the online supplemental material, we see that the maximum efficiency is achieved by sampling from U_v. At the same time, collecting the value of the variable L_{cj} could be complex due to the need to set up a *back-tracing process*. Thus, a feasible alternative strategy for the estimation of Y could be represented by:

$$\hat{Y}_{alt} = \hat{Y}_A + \hat{Y}_C - \hat{Y}^A_{AC},$$

where $\hat{Y}_C$ is the standard -HT estimate of the total y in U_C and $\hat{Y}^A_{AC}$ is the GWSM estimate of the total y in the intersection of U_A with U_C obtained by the sample S_A. These terms are as follows:

$$\hat{Y}_c = \sum_{k \in S_C} \frac{1}{\pi_{Ck}} y_k,$$

$$\hat{Y}^A_{AC} = \sum_{k \in S_v} \frac{1}{\pi_{vk}} \sum_{j \in S_{vk}} \frac{1}{\pi_{2v|k}} \frac{1}{L_{vj}} l_{k,j} y_j \mathbb{I}\left[\left(L_j - L_{Cj}\right) \geq 1\right],$$

where L_j is the total number of contacts of unit j.

6. Sampling Design for Follow-Ups of the Survey in Subsequent Waves

The observational scheme proposed in the above sections is set up as a cross-sectional survey. However, it can be adapted for monitoring the evolution of the number of infected people over time; this is done according to a mechanism that is updated like a chain mechanism time after time. While an in-depth study of this aspect deserves a separate study, we limit ourselves here to introducing the topic and to providing some initial indications.

Let us consider two consecutive points in time, for example, $t = 0$ and $t = 1$.

Assume person k is verified as infected at time 0 and is hence denoted as $v_{0,k} = 1$. This person may still be infected ($v_{1,k} = y_{1,k} = 1$), or she/he may no longer be infected ($y_{1,k} = 0$)

because of *death* (denoted by the dichotomous variable $d_{1,k} = 1$) or *healing* (denoted by the dichotomous variable $h_{1,k} = 1$).

The total of the y variable at time 1 may then be defined as:

$$Y_1 = Y_0 + \Delta D_{0\to 1} + \Delta H_{0\to 1} + \Delta Y_1, \tag{34}$$

where Y_0 is the total number of infections at time 0 and:

$$\Delta D_{0\to 1} = \sum_{k\in U} y_{0,k} d_{1,k}, \quad \Delta H_{0\to 1} = \sum_{k\in U} y_{0,k} h_{1,k}, \quad \Delta Y_1 = \sum_{k\in U} (1 - y_{0,k}) y_{1,k}. \tag{35}$$

In Equation (35), the quantity $(Y_0 + \Delta D_{0\to 1} + \Delta H_{0\to 1})$ indicates the total number of verified infected people at time 0 who are still infected at time 1, while the quantity ΔY_1, denotes the total number of *new* infections.

The updating of the sampling structures illustrated in the previous sections allows us to obtain a direct estimate of each of the components of Equation (34), as illustrated in Figure 2.

The total ΔY_1 can be estimated, as described in Section 5, using two sources of data, namely:

1. The sample $S_{1,v}$, which automatically captures the new entrants into the verified infected population at time 1. These new entrants are denoted by $\Delta U_{1,v}$, since the sampling selection is carried out on this population continuously over time. Then, a sample of their contacts can obtained as described in Subsection 4.2, resulting in the sample $S_{1,A}$.
2. The panel $S_{0,C}$, which is selected at time $t = 0$ and is updated over time, since the tests carried out at time $t = 1$ on the individuals of $S_{0,C}$ distinguish the *newly infected* people of the panel. Then, tracking the contacts of the infected people allows us to obtain the sample $S_{1,B}$.

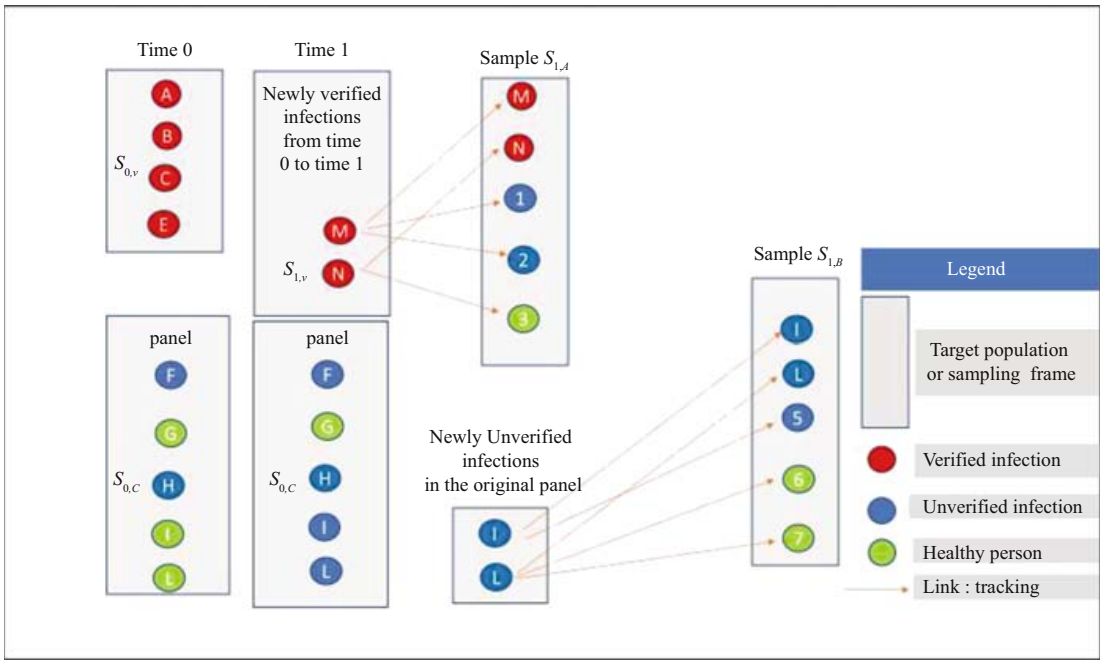

Fig. 2. Follow-up of samples over time.

The estimation of the totals $(Y_0 + \Delta D_{0\to 1} + \Delta H_{0\to 1})$ can be obtained by following up on the health statuses of the infected people captured by the samples $S_{0,A}$ and $S_{0,B}$ at time 0. The estimates are then obtained with the sampling weights computed at time 0.
Therefore, we have:

$$\hat{Y}_1 = \hat{Y}_0 + \widehat{\Delta D}_{0\to 1} + \widehat{\Delta H}_{0\to 1} + \widehat{\Delta Y}_1, \tag{36}$$

where $\hat{Y}_0, \widehat{\Delta D}_{0\to 1}, \widehat{\Delta H}_{0\to 1}, \widehat{\Delta Y}_1$ are the direct estimates of the quantities $Y_0, \Delta D_{0\to 1}, \Delta H_{0\to 1}, \Delta Y_1$, respectively. The above mechanism can be updated in a chain mechanism, and thus the estimate for any time $t > 1$ is obtained as:

$$\hat{Y}_t = \hat{Y}_{t-1} + \widehat{\Delta D}_{t-1\to 1} + \widehat{\Delta H}_{t-1\to t} + \widehat{\Delta Y}_t. \tag{37}$$

7. Empirical Evaluations of the Proposed Method: A Monte Carlo Study

7.1. The Design of Simulation Study

Since it is not possible at this stage to include a numerical illustration using real-life sample data, in this subsection, we report the results of a series of Monte Carlo experiments that numerically justify our proposed ideas and show their statistical performances in an artificial, although as realistic as possible, context.

Before showing our simulation results, we need to clarify the criteria we used in the data generation process and those employed in the generation of the geographical map on which the data are observed. This second element is essential given the peculiar nature of the transmission mechanism, which requires physical proximity between infected people. First, to simulate an artificial population describing the time evolution of an epidemic, we considered a popular model constituted by a system of six differential equations that, at each moment in time, describe six categories of individuals, namely, susceptible people (S), those exposed to the virus (E), those infected with symptoms (I), those without symptoms (A) and those that are removed from the population either because they healed (R) or are dead (D). This modelling framework is a result of the seminal contribution of Hamer (1906), Kermack and McKendrick (1927) and Soper (1929), and it is often referred to as the "SIR model" due to the initials of the categories considered in the first simplified formulation: Susceptibles, Infected and Removed. A comprehensive overview of this model is contained in Cliff et al. (1981). See also Vynnycky (2010).

Figure 3 diagrammatically describes the transitions between the six categories. For the random data generation process, we assumed that if infected, a susceptible person in the population (S) would remain in the exposed state (E) for five days. After that period, the subject could either become infected with symptoms (I) with probability 0.25 or without symptoms (asymptomatic; symbol A) with probability 0.75 (Bassi et al. 2020). An asymptomatic person remains infected (and so is still able to transmit the virus) for 14 days. After this period, all asymptomatic patients are considered healed and pass to the "removed" category (R). In contrast, the infected people showing symptoms heal with a probability of 0.85 or die (D) with a probability of 0.15 (death rate case).

For the map generation process, we considered a population distributed across 25 spatial units laid on a regular five-by-five square lattice grid. Each square of the grid contains a

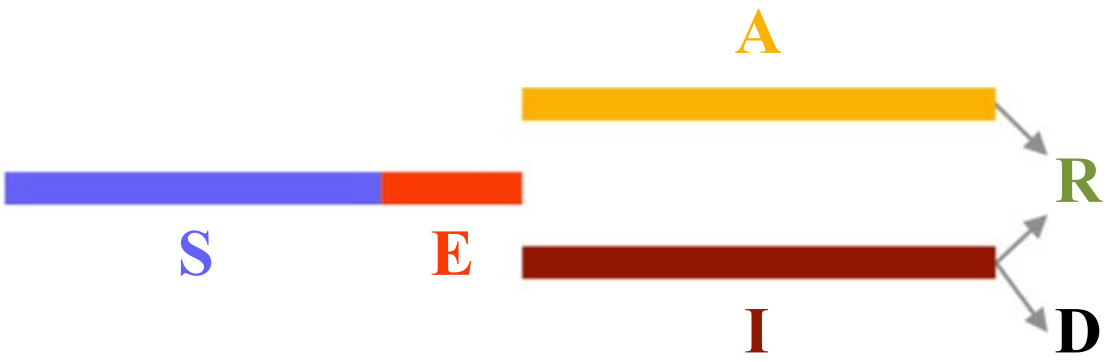

Fig. 3. The six basic categories of our simulation model and their transition patterns.

number of individuals randomly drawn from a uniform distribution ranging between 800 and 1,000. After performing a simulation exercise with these parameter values, we obtained an artificial population with a total of 22,217 individuals.

This geographical representation is very general in that the map generated in this way can represent, for example, a city divided into blocks, a region divided into smaller spatial unions, or any other meaningful geographical partition.

The contagion mechanism is favoured for studying human mobility. In our exercise, we assumed that at any moment in time, a certain percentage m of the population could move between the squares. We distinguished two epidemic phases. In Phase 1, people are free to move, and this percentage is $m = 0.03$, while Phase 2 describes a period of lockdown when mobility is discouraged and $m = 0.01$. In particular, we considered Phase 1 as a period of four weeks and Phase 2 as the period containing the eight subsequent weeks.

Communication during the lockdown period is limited not only by the number of people who move but also by the extent of their movements. This is a further simulation parameter that is generated by a uniform distribution ranging from -4 to 4 during Phase 1 (thus allowing movements in and out of the cells) and between -1 and 1 during Phase 2. Given the mobility pattern described above, contagion is determined by social interactions and contact opportunities. The number of contacts in each square of the grid is assumed to be determined by a random number drawn from a Poisson distribution with a parameter, that is, c_n, while the number of people involved in the movements is also a Poisson number characterised by a different parameter c_p. Given these assumptions, contagion occurs in the following way. If in a meeting at least one asymptomatic or exposed person is present, i_m susceptible people are infected and are moved into the "exposed" category. In our runs of the simulation, we considered Phase 1 to be characterised by the following parameters: $c_n = 20$; $c_p = 5$; $i_m = 3$. In contrast, during Phase 2, the three parameters became $c_n = 3$, $c_p = 3$, and $i_m = 2$ reflecting the decreased chances of contact between people. Figure 2 describes the time evolutions of the six categories of people in our simulated epidemics. As already stated, we considered Phase 1 to include four weeks (day 1 to day 28) and Phase 2 to include eight weeks (day 29 to day 84).

Figure 4 shows that despite the many assumptions that we were forced to include in the simulation, the contagion curves are very similar to those observed worldwide in the recent 2020 SARS-CoV-2 pandemic.

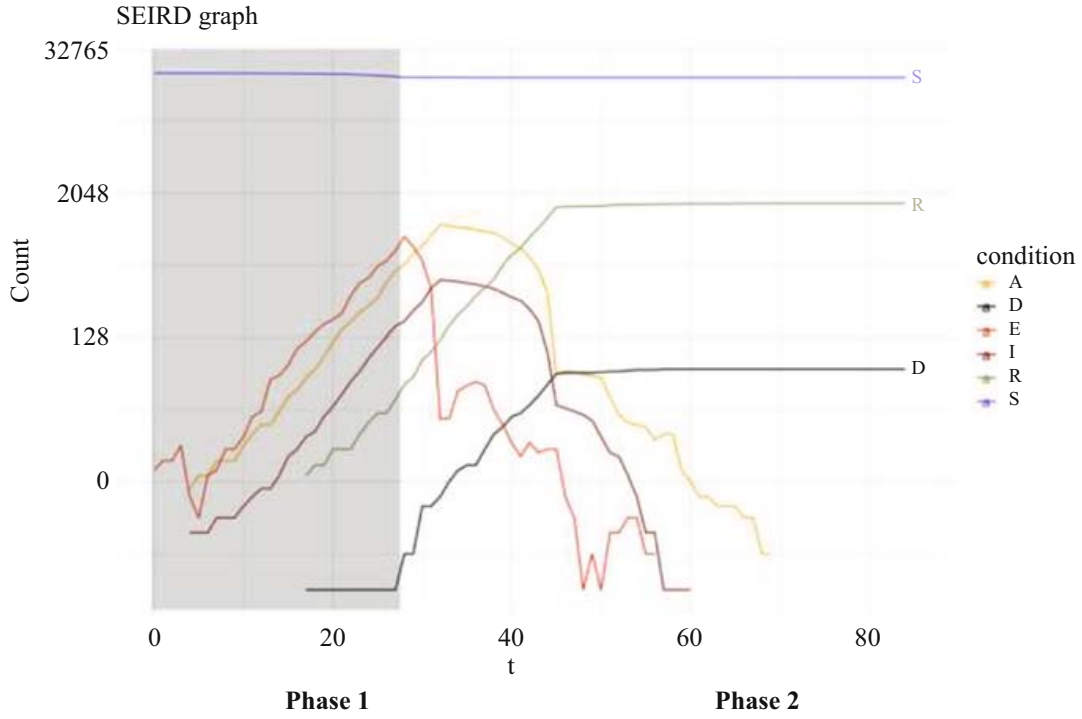

Fig. 4. Time evolutions of the six categories of people in the simulated epidemics. Phase 1 refers to days 1–28. Phase 2 refers to days 29–84.

7.2. Simulation Results

We present the main results obtained in the simulation exercise. Using the artificial population generated as described in the previous section, we considered the situation of a repeated sampling survey realized at three moments in time, namely, at day 15 (during the beginning of Phase 1), at day 25 (still in Phase 1, but in a situation closer to a *plateau*) and at day 35, during the period of lockdown. The infection situations at the three timepoints are reported in Table 1, where the results of the samples in Groups A and B and their intersection (see Figure 2) are separated. The sampling procedures are controlled by a set of parameters. For group A, the parameter g controls the contact tracing and represents the proportion of people sampled from all the contacts of verified infections. For group B, the parameter f is the proportion of healthy or unverified infected people sampled and v is the maximum number of contacts for each unit sampled. In Table 1, for group A we fixed the parameter $g = 0.9$, while for group B, the parameters f and v were fixed as follows: $f = 0.2$; $v = 12$.

Table 1. True simulated population values of the infected people for the two groups and their intersection observed on different days.

Groups	Day 15	Day 25	Day 35
Y_A	42	374	1,041
Y_B	126	875	1,432
Y_{AB}	39	372	1,018
Total infected	129	877	1,455

The sample sizes obtained with such parameter definitions (both excluding and including the contacts) are reported in Table 2 by distinguishing between four sample situations, namely, (1) A1B2: when both the individuals belonging to Group A and all their contacts ($g = 1$) are sampled while in Group B ($f = 0.2$) all contacts are sampled; (2) A1B3: when both the individuals belonging to Group A and their contacts are sampled ($g = 1$), while in Group B, the noninfected are sampled ($f = 0.2$) with all contacts (with a maximum of $v = 12$); (3) A2B2: when all individuals belonging to Group A but only a subset of their contacts are included in the sample ($g = 0.9$), and in Group B, the noninfected are sampled with all their contacts; and, finally, (4) A2B3: when all individuals belonging to Group A but only a subset of their contacts are included in the sample ($g = 0.9$), while in Group B, the noninfected are sampled with all their contacts but only up to a maximum of $v = 12$ individuals. Note that on day 35, we have fewer contacts in the sample than on day 25 due to the lockdown measures considered.

The main results of the simulation are reported in Table 3, and they show that in all sampling settings, the relative bias of our scheme is very small, and our estimators dramatically outperform simple random sampling in terms of efficiency (the ratio of the standard error of the proposed estimator computed by the simulation to that of the HT estimator for simple random sampling without replacement). In particular, the relative bias is on the order of 0.01% during Phase 1, while during Phase 2, it depends on the adoption of a sampling scheme with high precision when both the individuals belonging to Group A and their contacts are included in the sampling process. In contrast, the relative bias obviously increases when only a subset of the contacts is observed. Furthermore, our

Table 2. Total number of sampling units with and without contacts on different days and in the various sampling schemes.

Day	Proportion of infected people in the population	Sampling units without contacts	Sampling scheme	Sampling units with contacts
15	0.006	4,130	A1B2	4,741
			A1B3	4,736
			A2B2	4,741
			A2B3	4,736
25	0.042	4,198	A1B2	7,650
			A1B3	7,634
			A2B2	7,650
			A2B3	7,634
35	0.070	4,361	A1B2	7,545
			A1B3	7,514
			A2B2	7,545
			A2B3	7,514

Sampling scheme description: A1 = All individuals in Group A and their contacts are totally sampled; A2 = All individuals in Group A subset of their contacts are sampled; B2 = A subset of Group B and all their contacts are sampled; B3 = A subset of Group B and all their contacts are sampled, but only up to a maximum of v = 12 individuals.

Table 3. Results of the simulation study for the various sampling schemes on different days.

Days	Percentage of infected people in the population	True population value (see Table 1)	Sampling scheme	Estimated total number of infected people (average over 500 simulations)	α^*	Standard error	Coefficient of variation (%) $\frac{(7)}{(5)} \times 100$	Relative absolute bias $\frac{\lvert(3)-(5)\rvert}{(3)}$	Relative efficiency compared with that of the simple random sample without contacts	Relative efficiency compared with that of the simple random sample with contacts
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)[a]
15	0.0058	129	A1B2	128.99	0.00	0.01	0.09	0.0001	0.0006	0.0006
			A1B3	128.99	0.00	0.01	0.09	0.0001	0.0006	0.0006
			A2B2	128.75	0.00	1.56	0.97	0.0019	0.0659	0.0718
			A2B3	128.75	0.00	1.56	0.97	0.0019	0.0659	0.0718
25	0.0394	877	A1B2	876.83	0.00	0.17	0.05	0.0001	0.0027	0.0041
			A1B3	876.84	0.00	0.16	0.05	0.0001	0.0027	0.0040
			A2B2	876.90	1.00	0.48	0.08	0.0001	0.0080	0.0120
			A2B3	877.02	0.48	2.43	0.18	0.0000	0.0403	0.0605
35	0.0654	1.455	A1B2	1,455.00	0.00	0.00	0.00	0.0000	0.0000	0.0000
			A1B3	1,455.00	0.00	0.00	0.00	0.0000	0.0000	0.0000
			A2B2	1,461.59	0.00	9.78	0.21	0.0045	0.1310	0.1895
			A2B3	1,461.59	0.00	9.78	0.21	0.0045	0.1310	0.1895

Sampling scheme description: A1 = All individuals in Group A and their contacts are totally sampled; A2 = All individuals in Group A and bset of their contacts are sampled; B2 = A subset of Group B and all their contacts are sampled; B3 = A subset of Group B and all their contacts are sampled, but only up to a maximum of v = 12 individuals.
Columns (10) and (11): The relative efficiency is computed as the ratio of the standard error of the proposed estimator (computed by the simulation) to that of the HT estimator for simple random sampling without replacement.

Table 4. Comparison of the relative efficiency with different proportions for Group B on day 35.

Sampling scheme	$f = 0.2$	$f = 0.02$
A1B2	0,00	0,06
A1B3	0,00	0,06
A2B2	0,19	0,38
A2B3	0,19	0,38

method outperforms the simple random sample in terms of efficiency. Similar to the case of bias, the relative advantage of our scheme over the simple random sample with respect to efficiency is greatest in the case of the A1 sample scheme when all selected individuals and their contacts are included in the sample, while it is lowest in the case of A2 when only a subset of them is observed. Moreover, Table 3 also displays a decrease in the relative advantage of our method for the day 35 wave, where due to the lockdown restrictions, the number of contacts is very limited.

The results presented here depend greatly on the particular settings of the (many) parameters involved in the simulation that describe different epidemic evolutions. To mitigate such subjectivity, we also run many other Monte Carlo experiments using different parameter values. Although they are available upon request, these results are not reported here for the sake of succinctness. However, they all confirm the same features: our method has a very low relative absolute bias and it is superior to the simple random sampling scheme in terms of efficiency.

In order to measure the relative efficiency with different sample size, we compare the sampling schemes setting the proportion of healthy or unverified infected people sampled (f) as 0.2 and 0.02. The results in Table 4 confirm that, even if the ratio between sample size and population decreases, the efficiency of the methodology is confirmed.

8. Conclusions and Future Challenges

The aim of this article is to draw the attention of researchers and decision makers to the need to observe the characteristics of the COVID-19 pandemic through a formal sampling design, thus overcoming the limitations of data collected on a convenience basis. Only in this way will we be able to produce both reliable estimates of the current situation and forecasts of the future evolution processes of epidemics so that we can make empirically grounded decisions about public health monitoring and surveillance, especially in the transition phase between the decline from the epidemic peak and the relaxation of quarantine measures (Alleva 2017).

In such a situation, data must be comparable over time. It is essential to set up a system of data collection that allows for statistically valid comparisons over time and across different geographic areas by taking different economic, demographic, social, environmental and cultural contexts into account.

We believe that clear knowledge of the phenomenon is also necessary for the population to become aware of it and to adopt responsible behaviours. Trust and sharing must be grounded on a solid information base.

In comparison with other possible observational strategies, the proposal in this article has three elements of strength:

1. Relevance. The proposed sampling scheme, designed to capture most of the infected people through an indirect sampling mechanism, not only aims at providing a snapshot of the phenomenon at a single moment in time but is designed as a continuous survey that repeated in several waves over time. It also takes the different target variables in different stages of epidemic development into account and contributes to the implementation of a statistical surveillance system for the epidemic that could be integrated with existing systems managed by the health authorities.
2. Accuracy. In this article, the properties of the estimators have been formally proven and confirmed by analysing the results of a set of Monte Carlo experiments. The results guarantee the reliability of the estimators in terms of unbiasedness, and their efficiency is higher than that of a simple random sample.
3. Timeliness. The sampling design is operable immediately, as this is required by the emergency we are experiencing. Indeed, this article represents the statistical formalization of a recent proposal (Alleva et al. 2020) and has been accompanied by a technical note that describes the different phases into which it is divided, the subjects involved and the crucial aspects required for its success (Ascani 2020).

Relevance, accuracy and Timeliness are quality dimensions proposed by the European Statistical System (Eurostat 2017). Although our effort with regard to the pandemic has progressed during this phase of the emergency, there is room for much methodological and statistical research in terms of setting up statistical instruments for producing reliable and timely estimates of the phenomenon. Indeed, from a methodological point of view, while in this article we have fully derived the properties of the estimators in the cross-sectional case, the properties in subsequent waves still need to be proven formally. Among other aspects to be developed, we mention those related to time and spatial correlations, which are useful both for modelling the phenomenon and for designing an efficient spatial sampling technique to achieve the same level of precision as that of the current method but with fewer sample units (Arbia and Lafratta 2002). A specific extension of the spatial sampling techniques to be further developed is the use of capture/recapture techniques (Borchers 2009; Lavallée and Rivest 2012), which would require an overlap of the samples in Groups A and B. A further improvement to be explored could be derived from applying the Dorfmann procedure (Dorfmann 1943) to reduce the number of tests and the cost of our method.

In addition to the methodological advances, other general aspects to be developed in concert with different specialists are the integration of the statistical system proposed here with the health authority's surveillance system for infected people and the use of their contact- tracking devices for statistical purposes. These devices could be useful both for the identification of contacts and for measuring the propensity of people to travel and the connected risks of doing so. To this end, it could be interesting to study the possibility of considering, within our framework, the proposal developed by Saunders-Hastings et al. (2017), who addressed the problem of monitoring during a pandemic via a model approach. The need to monitor pandemics over time should

represent the motivation for building an integrated surveillance system. This system should merge three different pieces of information within a unified database: (1) the information collected by the administrative institutions when receiving and treating individuals who have turned to the healthcare system; (2) the statistical information collected on purpose with the aim of accurately measuring the diffusion of an infection; and finally, (3) the data obtained through new sources for tracking the movements of people and their contacts.

A third extension of our proposal concerns the operational point of view. Indeed, the sampling design described in detail in Section 5 should be accompanied by the definitions of some key points:

- A control room that ensures the necessary inter-institutional collaboration for guiding field operations (Health Authorities at the national and regional levels, Statistical Offices, others).
- An effective information campaign to promote participation among the population; the required legal framework to assure the collection and analysis of personal data.
- A medical testing procedure to consider for the selected population (swabs, blood testing and DNA mapping).
- The geographical-temporal estimation domains of interest and the sample dimensions on the basis of the information needs and the available financial and organizational resources.
- The frequency of sampling for Groups A and B, as well as the length of stay in the panel of group B.
- The sociodemographic characteristics, living conditions and mobility behaviours to be collected at the individual and family levels to shed light on relative risks and to evaluate the effects of the policies adopted for modifying the evolution of the epidemic.

This can only be achieved if epidemiologists, virologists, and administrators of healthcare institutions work in conjunction with experts in mathematical and statistical modelling and forecasting and experts in the evaluation of public policies.

We designed the sampling mechanism considering the Italian situation, and we proved its feasibility by defining the previous key points to estimate the times and costs of our method (Alleva et al. 2020). The sample size required to assure a certain level of accuracy for the estimates depends on the base rate infection. The unit cost of administering the swab and serological tests relies on the level of involvement in the survey by the public health authorities. The total cost depends on the length and the periodicity of the panel survey. For Italy, we estimated the cost of data collection at the national and regional levels (21 regions), for a case with three months of monitoring, a panel survey every 15 days and a base rate of infection of 0.04. With regard to Groups A and B, the sample sizes are 1,000 and 1,200 units, and this implies requirements of 6,000 and 7,200 swabs and total costs of 210,000 and 252,000 euros, respectively. In adopting the suggested strategy, different countries may require adjustments to take the peculiarities of their specific health system and institutional framework into account. For this research direction, the contributions of the National Statistical Offices, as well as common actions and the sharing of experiences at the European and worldwide levels, will be essential.

9. References

ABC. 2020. "Random coronavirus testing to begin in Canberra next week at drive-through centre and clinic". *ABC News*. Available at: https://www.abc.net.au/news/2020-04-03/random-coronavirus-testing-begins-in-canberra/12119364 (accessed April 2020).

Aguilar, J.B., J.S. Faust, L.M. Westafer, and J.B. Gutierrez. 2020. "Investigating the Impact of Asymptomatic Carriers on C0VID-19", *medXiv*. DOI: https://doi.org/10.1101/2020.03.18.20037994.

Alleva, G. 2017. "The new role of sample surveys in official statistics". ITACOSM 2017, The 5th Italian Conference on Survey Methodology, June 14, 2017. Bologna IT. Available at: https://www.istatit/it/files//2015/10/Alleva_ITAC0SM_14062017.pdf (accessed April 2020).

Alleva G. 2020. *Contributo per la 12° Commissione permanente Igiene e sanità del Senato della Repubblica.* May 27, 2020. Roma, IT. https://www.senato.it/application/xmanager/projects/leg18/attachments/documento_evento_procedura_commissione/files/000/135/501/GI0RGI0_ALLEVA.pdf (accessed May 2020).

Alleva, G., G. Arbia, P.D. Falorsi, G. Pellegrini, and A. Zuliani. 2020. *A sampling design for reliable estimates of the SARS-CoV-2 epidemic's parameters. Calling for a protocol using panel data*. https://web.uniroma1.it/memotef/sites/default/files/Proposal.pdf (accessed April 2020).

Alleva G., and A. Zuliani. 2020. "Coronavirus: chiarezza sui dati", *Bancaria*. Available at: https://www.bancaria.it/livello-2/archivio-sommari/gli-ultimi-sommari-di-bancaria/-bancaria-giugno-2020/covid-19-chiarezza-sui-dati/.

Arbia, G. 1994. "Selection techniques in sampling spatial units", *Quaderni di statistica e matematica applicata alle scienze economico-sociali*, XVI(1–2): 81–91.

Arbia, G. 2020. *A note on early epidemiological analysis of coronavirus disease 2019 outbreak using crowdsourced data.* arXiv.

Arbia, G. and G. Lafratta. 1997. "Evaluating and updating the sample design: the case of the concentration of SO_2 in Padua", *Journal of Agricultural, Biological and Environmental Statistics*, 2, 4: 451–466. DOI: https://doi.org/10.2307/1400514.

Arbia, G., and G. Lafratta. 2002. "Spatial sampling designs optimized under anisotropic superpopulation models", *Journal of the Royal Statistical Society series c – Applied Statistics*, 51, 2: 223–23.

Ascani, P. 2020. *Technical Note on the methods of the data collection phase for a proposal of sampling design for reliable estimates of the epidemic's parameters of SARS-CoV-2.* Available at: https://web.uniroma1.it/memotef/sites/default/files/TechNote.pdf (accessed May 2020).

Bassi F., G. Arbia, and P.D. Falorsi. 2020. "Observed and estimated prevalence of Covid-19 in Italy: How to estimate the total cases from medical swabs data, from medical sbabs data". *Science of the Total Environment*, 764: 142799. DOI: https://doi.org/10.1016/j.scitotenv.2020.142799.

Borchers, D. 2009. "A non-technical overview of spatially explicit capture-recapture models". *Journal of Ornithology*, 152: 435–444. DOI: https://doi.org/10.1007/s10336-010-0583-z.

Chughtai, A.A., and A.A. Malik. 2020. "Is Coronavirus disease (C0VID-19) case fatality ratio underestimated?". *Global Biosecurity*, 1(3). DOI: http://doi.org/10.31646/gbio.56.

Cliff, A.D., P. Haggett, J.K. Ord, and F.R. Verfey. 1981. *Spatial Diffusion: an Historical Geography of Epidemics in an Island Community* 14. Cambridge University Press.

Cochran, W.G. 1977. *Sampling Techniques*. Wiley. New York.

Deville, J.-C., and Y. Tillé. 2004. "Efficient Balanced Sampling: the Cube Method", *Biometrika* 91: 893–912. DOI: https://doi.org/10.1093/biomet/91.4.893.

Deville, J.-C., and Y. Tillé. 2005. "Variance approximation under balanced sampling", *Journal of Statistical Planning and Inference* 128: 569–591. DOI: https://doi.org/ 10.1016Zj.jspi.2003.11.011.

Dewatripont, M., M. Goldman, E. Muraille, and J.-P. Platteau. 2020. "Rapidly identifying workers who are immune to COVID-19 and virus-free is a priority for restarting the economy", *VoxEU*. Available at: https://voxeu.org/article/rapidly-identifying-workers-who-are-immune-covid-19-and-virus-free-priority-restarting-economy (accessed March 2020).

Di Gennaro Splendore, L. 2020. "Random testing, quality of data and lack of information: COVID-19". *Data and Policy*. Available at: https://medium.com/data-policy/random-testing-quality-of-data-and-lack-of-information-covid-19-a6e09a398d1d (accessed April 2020).

Dorfman, R. 1943. The Detection of Defective Members of Large Populations, *The Annals of Mathematical Statistics* 14 (4): 436–440. DOI: http://dx.doi.org/10.1214/aoms/1177731363.

Eurostat. 2017. *European statistics Code of Practice – revised edition.* Available at: https://ec.europa.eu/eurostat/web/products-catalogues/-/KS-02-18-142 (accessed May 2020)

Falorsi P.D., and P. Righi. 2015. "Generalized framework for defining the optimal inclusion probabilities of one-stage sampling designs for multivariate and multi-domain surveys". *Survey methodology* 41: 215–236. Available at: https://www150.statcan.gc.-ca/n1/en/catalogue/12-001-X201500114149.

Falorsi P.D., P. Righi, and P. Lavallée. 2019. "Cost optimal sampling for the integrated observation of different populations". *Survey methodology* 45(3): 485–511. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/12-001-X201900300004.

Ferretti, L, C. Wymant, M. Kendall, L. Zhao, A. Nurtay, L. Abeler-Dorner, M. Parker, D. Bonsall, and C. Fraser. 2020. "Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing", *Science*. DOI: http://dx.doi.org/10.1126/science.abb6936.

Fuggetta M. 2020. "Testing for the Base Rate". *Bayes*. Available at: http://massimofuggetta.com/2020/04/28/testing-for-the-base-rate/ (accessed April 2020).

Gros, D. 2020. "Creating an EU 'Corona Panel': Standardised European sample tests to uncover the true spread of the coronavirus". *VoxEU*. Available at: https://voxeu.org/article/standardised-european-sample-tests-uncover-true-spread-coronavirus (accessed May 2020).

Hackenbroch, V. 2020. "Groβe Antikörperstudie soll Immunität der Deutschen gegen Covid-19 feststellen" *Spiegel*. Available at: https://www.spiegel.de/wissenschaft/medizin/coronavirus-grosse-antikoerper-studie-soll-immunitaet-der-deutschen-feststel-

len-a-c8c64a33-5c0f-4630-bd73-48c17c1bad23?d = 1585300132&sara_ecid = soci_upd_wbMbjhOSvViISjc8RPU89NcCvtlFcJ. (accessed May 2020).

Hamer W.H. 1906. "Epidemic diseases in England", *Lancet*, 1. DOI: https://doi.org/10.1016/S0140-6736(01)80187-2.

Hansen N.H., N.W. Hurwitz, and W.G. Meadow. 1953. *Sample Survey Method and Theory*. Wiley, New York.

Hartley, H.O. 1962. "Multiple Frame Surveys", Proceedings of the Social Statistics Section, American Statistical Association, Alexandria, Va. 1962.

Hartley, H.O. 1974. "Multiple Frame Methodology and Selected Applications", *Sankhya*, 36: 99–118.

Horvitz, D.G., and D.L. Thompson. 1952. "A generalisation of sampling without replacement from finite-universe". *J Amer Statist. Assoc*. 47: 663–685. DOI: http://doi.org/0.1080/01621459.1952.10483446.

ILO (International Labour Organization). 2020. "COVID-19 impact on the collection of labour market statistics". https://ilostat.ilo.org/topics/covid-19/covid-19-impact-on-labour-market-statistics/ (accessed May 2020).

Ioannidis, J. 2020. "A fiasco in the making? As the coronavirus pandemic takes hold, we are making decisions without reliable data". *Statnews*. https://www.statnews.com/2020/03/17/afiasco-in-the-making-as-the-coronavirus-pandemic-takes-hold-we-are-making-decisions-withoutreliable-data/ (accessed Mar 2020).

Istat (Istituto nazionale di statistica). 2020. "Primi risultati dell'indagine di sieroprevalenza sul SARS-CoV-2. https://www.istat.it/it/files//2020/08/ReportPrimiRisultatiIndagineSiero.pdf (accessed Aug 2020).

Kermack, W.O., and A.G. McKendrick. 1927. "A contributions to the mathematical theory of epidemics" Proceedings of the Royal society London 115: 700–721.

Kiesl, H. 2016. "Indirect Sampling: A Review of Theory and Recent Applications". *AStA Wirtschafts und Sozialstatistisches Archiv*. DOI: http://doi.org/10.10.1007/s11943-016-0183-3.

Kish, L. 1965. *Survey Sampling*, Wiley. New York.

Lavallée, P. 2007. *Indirect Sampling*, Springer series in statistics.

Lavallée, P., and L.P. Rivest. 2012. "Capture-Recapture Sampling and Indirect Sampling". *Journal of Official Statistics* 28(1): 1–27. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/capture150recapture-sampling-and-indirect-sampling.pdf. (accessed March 2022).

Leung, G., and K. Leung. 2020. "Crowdsourcing data to mitigate epidemics, the lancet digital health", *The Lancet Digital Health*. DOI: https://doi.org/10.1016/S2589-7500(20)30055-8.

Li, R., S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman. 2020. "Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (SARS-CoV2)", *Science* 368 (6490): 489–493. DOI: http://doi.org/10.1126/science.abb3221.

Mizumoto, K., K. Kagaya, A., Zarebski, and G. Chowell. 2020. "Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess cruise ship, Yokohama, Japan, 2020". *Euro surveillance: bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin, 25*(10), 2000180. DOI: https://doi.org/10.2807/1560-7917.ES.2020.25.10.2000180.

Murthy M.N., and V.K. Sethi. 1965. "Self-Weighting Design at Tabulation Stage" *Sankhya: The Indian Journal of Statistics*, 27(1–2): 201–210.

Narain, R.D. 1951. "On sampling without replacement with varying probabilities". *Journal of the Indian Society of Agricultural Statistics* 3: 169–174.

ONS (Office for national Statistics). 2020. *Coronavirus (COVID-19) Infection Survey pilot: England and Wales.* Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/healthandsocialcare/conditionsanddiseases/bulletins/coronaviruscovid19infectionsurveypilot/englandandwales14august2020. (accessed August 2020).

Romania-insider.com. 2020. "Coronavirus in Romania: Over 10,000 Bucharest residents will be tested for Covid-19 as part of a study". *Romania-insider.com*. Available at: https://www.romania-insider.com/coronavirus-romania-bucharest-testing-streinu-cercel. (accessed April 2020).

Rossman, H., A. Keshet, S. Shilo, A. Gavrieli, T. Bauman, O. Cohen, R. Balicer, B. Geiger, Y. Dor, and E. Segal. 2020. "A framework for identifying regional outbreak and spread of COVID-19 from one- minute population-wide surveys". *Nature Medicine* 26(5): 634–638. DOI: https://doi.org/10.1101/2020.03.19.20038844.

Saunders-Hastings, P., B.Q. Quinn Hayes, R. Smith, and D. Krewski. 2017. "Control strategies to protect hospital resources during an influenza pandemic". *PloS one* 12(6): e0179315. DOI: https://journals.plos.org/plosone/article?id = 10.1371/journal.pone.0179315.

Scott J. 2000. *Social Network Analysis. A Handbook, London*, Sage Publications.

Singh, A.C., and F. Mecatti. 2011. "Generalized Multiplicity-Adjusted Horvitz-Thompson Estimation as a Unified Approach to Multiple Frame Surveys". *Journal of Official Statistics* 27(4): 633–650. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/generalized-multiplicity-adjusted-horvitz-thompson-estimation-as-a-unified-approach-to-multiple-frame-surveys.pdf (accessed Marsch 2022).

Soper H.E. 1929. "Interpretation of periodicity in disease prevalence", *Journal of the Royal Statistical Society* A 92: 34–73. DOI: https://doi.org/10.2307/2341437.

Sudman, S., G. Monroe, M.G. Sirken, and C.D. Cowan. 1988. "Sampling Rare and Elusive Populations" *Science* 240(4855): 991–996. Available at: https://www.science.org/doi/10.1126/science.240.4855.991.

Sun., K., J. Chen, and C. Viboud. 2020. "Early epidemiological analysis of coronavirus disease 2019 outbreak using crowdsourced data: a population level observational study", *The Lancet Digital Health*. DOI: https://doi.org/10.1016/S2589-7500(20) 30026-1.

Tillé Y., and M. Wilhelm. 2017. "Probability Sampling Designs: Principles for Choice of Design and Balancing". *Statistical Science* 32(2): 176–189. DOI: https://doi.org/10.1214/16-STS606.

Thompson S.K., and G.A.F. Seber. 1996. *Adaptive Sampling*. Wiley Series in Probability and Statistics, New York.

Vynnycky, E. 2010. *An Introduction to Infectious Disease Modelling*, edited by R.G. White. Oxford: Oxford University Press.

Yang S., F.B. Keller, and L. Zheng. 2016. *Social Network Analysis: Methods and Examples*, Sage Publications, London.

Yelin, I., N. Aharony, E. Shaer-Tamar, A. Argoetti, E. Messer, D. Berenbaum, E. Shafran, A. Kuzli., N. Gandali, T. Hashimshony, Y. Mandel-Gutfreund, M. Halberthal, Y. Geffen,

M. Szwarcwort-Cohen, and R. Kishony. 2020. "Evaluation of C0VID-19 RT-qPCR test in multi-sample pools", *medRxiv*. DOI: https://doi.org/10.1101/2020.03.26. 20039438.

Received April 2020
Revised August 2020
Accepted March 2021

Chapter 3

Processing

Effects of Confidentiality-Preserving Geo-Masking on the Estimation of Semivariogram and of the Kriging Variance

geographical analysis

Check for updates

Geographical Analysis (2023) **55**, 466–481

Effects of Confidentiality-Preserving Geo-Masking on the Estimation of Semivariogram and of the Kriging Variance

Giuseppe Arbia[1], Chiara Ghiringhelli[1], and Vincenzo Nardelli[2]

[1]Dipartimento di Scienze Statistiche, Università Cattolica del Sacro Cuore, Rome, Italy, [2]Department of Economics, Management and Statistics (DEMS), Università di Milano Bicocca, Milan, Italy

Geostatistical methods, such as semivariograms and kriging are well-known spatial tools commonly employed in many disciplines such as health, mining, forestry, meteorology to name only few. They are based essentially on point-referenced data on a continuous space and on the calculation of distances between them. In many practical instances, however, the exact point location, even if exactly known, is geo-masked to preserve confidentiality. This typically happens when dealing with confidential data related to individuals-health and their biometric parameters. In these situations, the estimation of the semivariogram and, hence, the spatial prediction can become biased and highly inefficient. This paper examines the extent of the bias in the particular case when the geo-masking mechanism is known (called "intentional locational error") and lays the ground to a full understanding of the phenomenon in more general cases. We also examine how the geo-masking affects the estimation of the kriging variance thus reducing the efficiency of spatial prediction. We pursue our aims by developing some theoretical results and by making use of simulated and real data analysis.

Introduction

Geostatistical methods, such as semivariograms and kriging, are well known spatial tools commonly employed in many disciplines such as health, mining, forestry, meteorology to name only few (Banerjee, Carlin, and Gelfand 2004; Shabenberger and Gotway 2005; Diggle and Ribeiro 2007; Montero, Fernàndez-Avilés, and Mateu 2015). The essence of the methods is to study the regularities observed on point-referenced data on a continuous two- (or three-) dimensional space. Such regularities are explored by modeling the differences between sample attribute pairs as function of distance. In many practical instances, however, the exact individuals' point location is not exactly known due to either survey imperfections or geo-masking intentionally introduced in order to preserve confidentiality (see e. g. Gabrosek and Cressie 2002 and Cressie and Kornak 2003). The interest in this paper is limited to what Arbia, Espa, and Giuliani (2016)

Correspondence: Giuseppe Arbia, Dipartimento di Scienze Statistiche, Università Cattolica del Sacro Cuore, Largo Francesco Vito 1, 00168 Rome, Italy
e-mail: giuseppe.arbia@unicatt.it

Submitted: November 10, 2020. Revised version accepted: July 29, 2022.

doi: 10.1111/gean.12344

call *intentional locational error* and the masking mechanism is perfectly known, while we do not consider the related subject of when the locational uncertainty is *unintentionally* present in the data set like in the case of area-to-point spatial interpolation (see Kyriakidis and Yoo 2005; Kyriakidis 2010).

One example of intentional locational error refers to health studies where, for instance, mortality maps allow the identification of spatial patterns, clusters and disease hot spots that can often inform the etiology of the phenomenon. Health data are often available at the individual level and the position of the single observational unit is frequently geo-referenced with a high level of accuracy using GPS. However, for the obvious scope of preserving the respondent's confidentiality, the individual's coordinates are often displaced before being disseminated (Allhouse et al. 2010; Seidl, Jankowski, and Clarke 2018). Understanding the distorting effects of geo-masking is crucial when individual geo-masked data are used to produce spatially interpolated maps related to phenomena of public health concerns. Good examples are provided by the Demographic and Health Surveys project run by the U.S. Agency for International Development (USAID) where individual data are collected on fertility, maternal and child health, diseases diffusion, malnutrition, and many others (Burgert et al. 2013). A sound use of geostatistical tools is also necessary to predict the likely health risk in unsampled locations (e. g. Webster et al. 1994; Oliver et al. 1998), or outpatient treatment burdens (Gething et al. 2005). Another important use of geographical individual data concerns mapping some of the crucial variables during an epidemic virus diffusion. For instance, in the Covid-19 pandemics, geographical interpolated maps of the *viral load* (i.e., the number of viral particles present in an infected individual) are of crucial importance due to the positive correlation with the severity of the illness (CEBM 2020) and to monitor the geographical diffusion of virus variants (Singanayagam et al., 2022).

For further examples of applications of geo-masking in geostatistical analysis of public health and for a review on the state of the art we refer to Goovaerts (2008) and Diggle and Giorgi (2021).

A second example concerns environmental data. For instance, forest inventories are used to monitor the state of the environment and to measure both tree's characteristics (such as biomass and growth) and other environmental variables (such as pollutant concentration) and to forecast them in unobserved (and often unobservable) locations (Zawadzki et al. 2005). Some of these analyses involve collecting data about trees whose position is geo-masked to preserve the information about the value of the trees and of their property and to avoid conflicts with the owners (Mcroberts et al. 2005; IFNC 2015).

Further situations may arise when using geostatistical methods in fields like social surveys (Grosh and Munoz 1996), social network analysis (Gao et al. 2019), social data (Pawitan and Steel 2006), crime (Kerry et al. 2010), and many others. When data are geo-masked, the use of geostatistical methods may lead to biased and inefficient predictions. In the literature, we find several approaches to handle the error introduced in the location, among the others we recommend Gabrosek and Cressie (2002), Cressie and Kornak (2003) and Fronterrè, Giorgi, and Diggle (2018).

In many situations, the geo-masking procedure is known. This is the case, for example, of the *random direction-random distance* method (see Collins 2011). Other geo-masking techniques were also proposed, but are less commonly employed in the literature (Cassa et al. 2006; Hampton et al. 2010; Zhang et al. 2015). See Zandbergen (2014) for a review.

Other contributes to positional error study are Santos et al. (2017) and Zhang and Roger (2000).

This paper aims at shedding light on the effects of geo-masking on geostatistical methods when the masking procedure is known. Our aim is to make researchers aware of the possible consequences of the presence of locational error while running empirical analyses and spatial prediction using geostatistical techniques. The paper is organized as follows. In Section 2 we examine the theoretical effects of locational errors on semivariogram estimation when the geo-masking mechanism is known. Section 3 examines some real data in the light of the previous results. Section 5 is devoted to examining how the geo-masking affects the kriging variance reducing the efficiency of the prediction. Section 4 concludes and describes possible future developments of the approach presented in this paper.

Effects of geo-masking on semivariogram estimation

Let us start considering, for the sake of exemplification and without loss of generality, an isotropic Gaussian covariance function (Banerjee, Carlin, and Gelfand 2004; Shabenberger and Gotway 2005) defined as:

$$c(d) = \sigma^2 \exp\left\{-3\frac{d^2}{\alpha^2}\right\} + \tau^2, \qquad \text{if } d \geq 0, \quad 0 \text{ otherwise}, \tag{1}$$

with d the pairwise distance between points, σ^2 the partial sill, τ^2 the nugget effect and α the effective range, that is the distance at which correlation decreases to less than 0.05. The inverse of α, say ϕ, is known as the decay parameter that controls the rapidity with which the covariance declines increasing the distance. The normalization factor of 3 is not essential, but is rather common in geostatistics.

The semivariogram associated to Equation (1) can be expressed as:

$$\gamma(d) = \sigma^2\left[1 - \exp\left\{-3\frac{d^2}{\alpha^2}\right\}\right] + \tau^2. \tag{2}$$

Both expressions (1) and (2) are function of the inter-point distance between two generic points of coordinates (x_i,y_i) and (x_j,y_j), say $d_{i,j}$, which is indicated as d, for short.

Let us now consider the case when the true coordinates of the points are intentionally masked to preserve confidentiality. In particular, let us consider the case when the coordinates are displaced according to a random mechanism such as the *random-direction random-distance* geo-masking procedure as it frequently happens in health surveys and in other situations (see Grosh and Munoz 1996; Mcroberts et al. 2005; Collins 2011; Burgert et al. 2013).

In this case a point observed in location (x_i,y_i) is randomly relocated within a circle of random radius θ_i and on a random angle δ_i such that $\theta_i \approx U(0,\theta^*)$, $\delta_i \approx U\left(0{,}360^\circ\right)$ with θ_i and δ_i mutually independent and identically distributed, and θ^* representing the maximum displacement distance. Let us call $\overline{d}$ the inter-point distance observed after geo-masking. In the Appendix we prove that:

$$E\left(\overline{d}^2\right) = d^2 + \frac{2}{3}\theta^{*2}. \tag{3}$$

Hence the geo-masking procedure introduces an expected upward bias in the estimation of the true pairwise distances which is quantified by the following expression:

$$E\left(d^2 - \overline{d}^2\right) = -\frac{2}{3}\left(\theta^*\right)^2.$$

In what follows we will assume that the original and geo-masked empirical semivariograms are the same and that the geo-masking only affects the parameter values and not the form of the semivariogram. Let us now consider, just for the sake of exemplification, the Gaussian covariogram with $\tau^2 = 0$ and, given the formulation adopted, let us measure the relative bias in the estimation of the covariance function at each distance d as follows:

$$B(d) = \ln\left(\frac{c(d)}{c(\bar{d})}\right). \tag{4}$$

In this case we have:

$$B(d) = \ln\left(\frac{\sigma^2 \exp\left\{-3\frac{d^2}{\alpha^2}\right\}}{\sigma^2 \exp\left\{-3\frac{\bar{d}^2}{\alpha^2}\right\}}\right) = -\frac{3}{\alpha^2}\left(d^2 - \bar{d}^2\right). \tag{5}$$

and thus, using Equation (3):

$$E[B(d)] = -\frac{3}{\alpha^2}E\left(d^2 - \bar{d}^2\right) = \frac{2}{\alpha^2}\left(\theta^*\right)^2. \tag{6}$$

Equation (6) shows that, in our hypotheses, when $\tau^2 = 0$, the expected bias introduced by a geo-masking procedure in the estimation of the covariogram at each distance d, is always positive, so that the covariogram is underestimated and hence the semivariogram is overestimated. Secondly, Expression (6) also shows that the expected bias is unaffected by the partial sill σ^2 while it depends proportionally on θ^* and inversely on the empirical range α. In particular, following intuition, $E[B(d)]$ increases with θ^*: the higher is the maximum displacement distance of the geo-masking, the larger is the bias. Conversely, the bias decreases with the empirical range α (so it increases with ϕ). In this case, indeed, if at low distances the covariogram flattens rapidly (when α is low so that there is a strong correlation at low distances) the bias is expected to be more severe. Conversely, if the covariogram stabilizes only at high distances (when α is high), the local covariance will be lower and the expected bias will be moderated. If a semivariogram is estimated with zero nugget effect, the theoretical overestimation is always confirmed. However, even if the nugget effect is not zero (as it happens in most empirical cases), an overestimation is always expected because the sign of the expected value of the bias in Equation (6) does not change if a positive constant is added to both the numerator and the denominator in the RHS of Equation (5).

Figs. 1 and 2 report a comparison between the true and of the error-infected isotropic Gaussian semivariogram in different artificial examples. Points are assumed to be laid on a unitary square, so that the theoretical maximum displacement distance is $\theta^* = \sqrt{2}$ corresponding to the square's diagonal. In our examples we considered different levels of geo-masking (ranging from moderate displacements, $\theta^* = 0.1$, corresponding to 7% of the maximum distance, to strong displacement, $\theta^* = 0.5$, corresponding to 35% of the maximum distance), and different empirical ranges including a sharp increase (high local correlation, $\alpha = 0.1$) and slow increase (low local correlation, $\alpha = 0.4$). Fig. 1 shows that the bias is higher in the case of low $\alpha = 0.1$ and moderated by $\alpha = 0.4$. All other things being equal the effect is more dramatic for high θ^*. When $\theta^* = 0.1$ and $\alpha = 0.4$ the semivariogram is substantially unaffected, but if α goes to 0.1 (implying a more rapid decrease in the covariance function) or θ^* goes up to 0.4, the effect clearly becomes more severe.

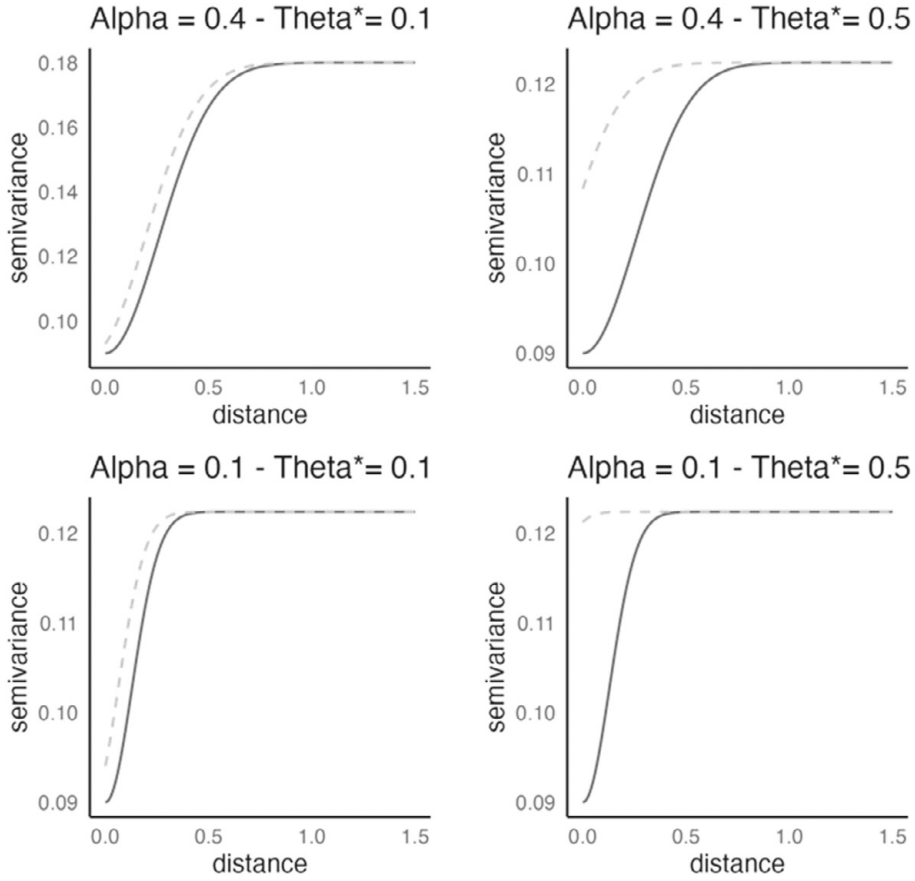

Figure 1. Semivariogram. Dashed light gray line is the contaminated semivariogram, bold black line is the true semivariogram. In all experiments $\tau = 0.3$; $\sigma = 0.3$; sill $= \tau^2 + \sigma^2 = 0.18$, nugget $= \tau^2 = 0.09$.

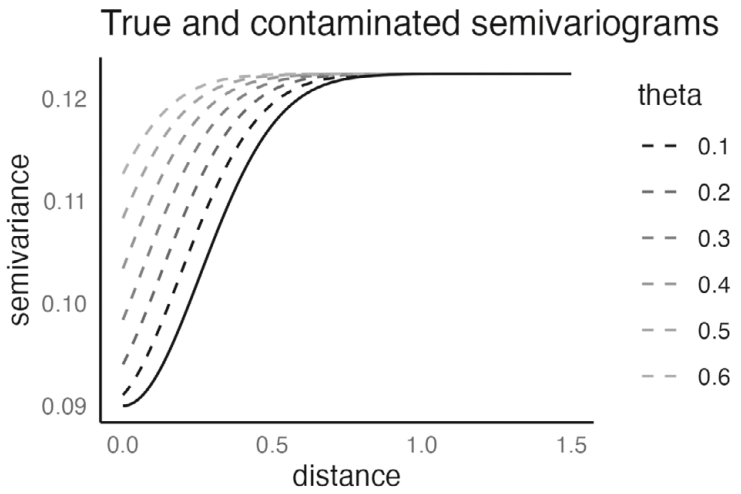

Figure 2. Effects of geo-masking on Gaussian semivariogram. Bold black line: True semivariogram. Dashed grays lines: Refer to increasing level of $\theta^* = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6$. In all cases $\alpha = 0.4$, $\sigma^2 = 0.18$, $\tau^2 = 0.09$.

To isolate the effect of θ^* on the estimation bias, Fig. 2 compares the true to the estimated covariance at a given level of $\alpha = 0.4$ considering different levels of $\theta^* = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6$. These levels correspond to a percentage of the maximum displacement 7, 14, 21, 28 35 and 42%. For high levels of $\theta^* > 0.3$, the geo-masking procedure produces a dramatic effect on the estimation of the semivariogram.

Real data analysis

The previous findings can be reinforced by examining a real data case. In what follows we will consider the data set `Meuse` available in the R library `gstat` (All R codes used in this paper are published in the repository https://github.com/vincnardelli/geomasking_kriging) and related to the quantity of four heavy metals measured along the river Meuse observed in 155 locations together with their spatial coordinates on a square of approximately 15-by-15 m^2 (Burrough and McDonnell 1998). For the sake of exemplification, Fig. 3 reports the map of the zinc quantity observed along the river.

The semivariograms estimated on the real data of the zinc quantity and the semivariograms estimated after geo-masking are reported in Fig. 4 for various level of the geo-masking parameter θ^*. In this section, to facilitate the comparison with the result of the preceding analysis based on the artificial data laid on a unitary square, we defined θ^* as a proportion of the maximum displacement distance given by the diagonal of the squared map on which data are laid.

The bias is evident already at very low values of the displacement parameter and, in particular, as soon as $\theta^* > 0.01$. Furthermore, as it was predicted by our theoretical analysis, apart from what is observed at the lower distances, the semivariogram is overestimated as it appears graphically when $\theta^* > 0.05$.

The results displayed in Fig. 4 could appear in contrast with those reported in Fig. 1 when $\boldsymbol{\theta^* = 0.1}$ and $\boldsymbol{\theta^* = 0.5}$ in that at the lower distances they report an underestimation rather

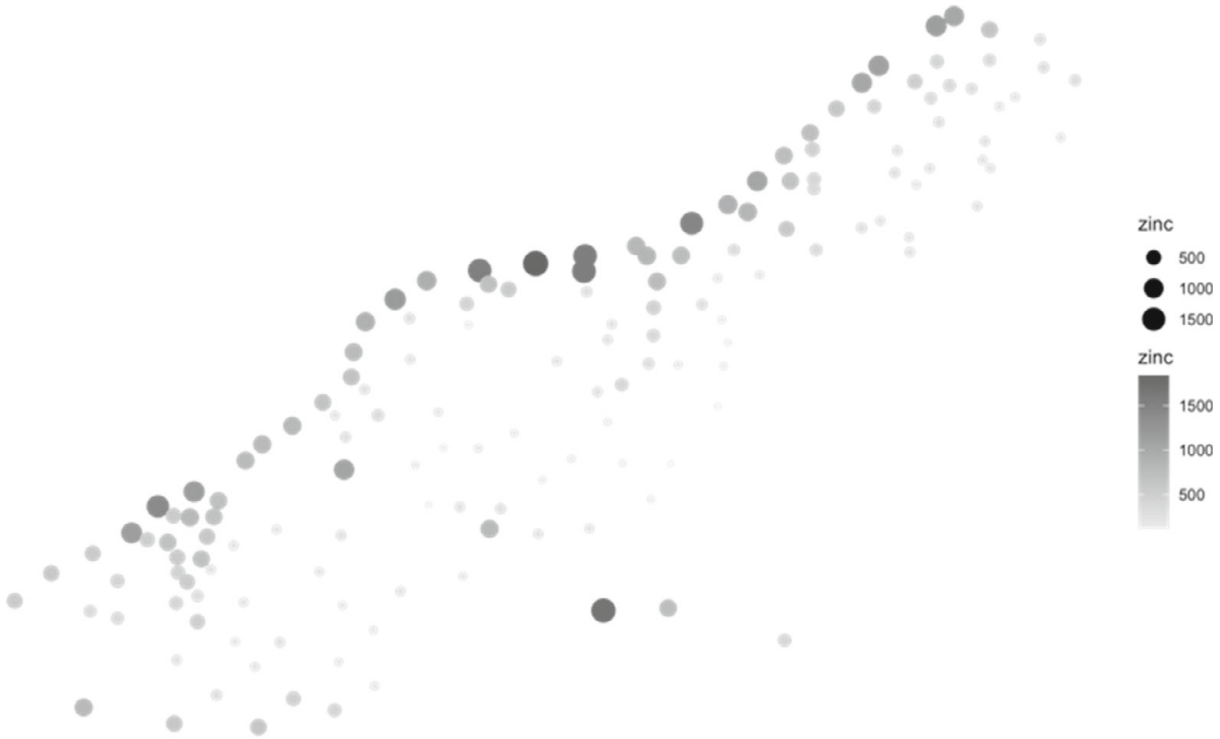

Figure 3. Map of the zinc quantity from the Meuse R data set in the library gstat.

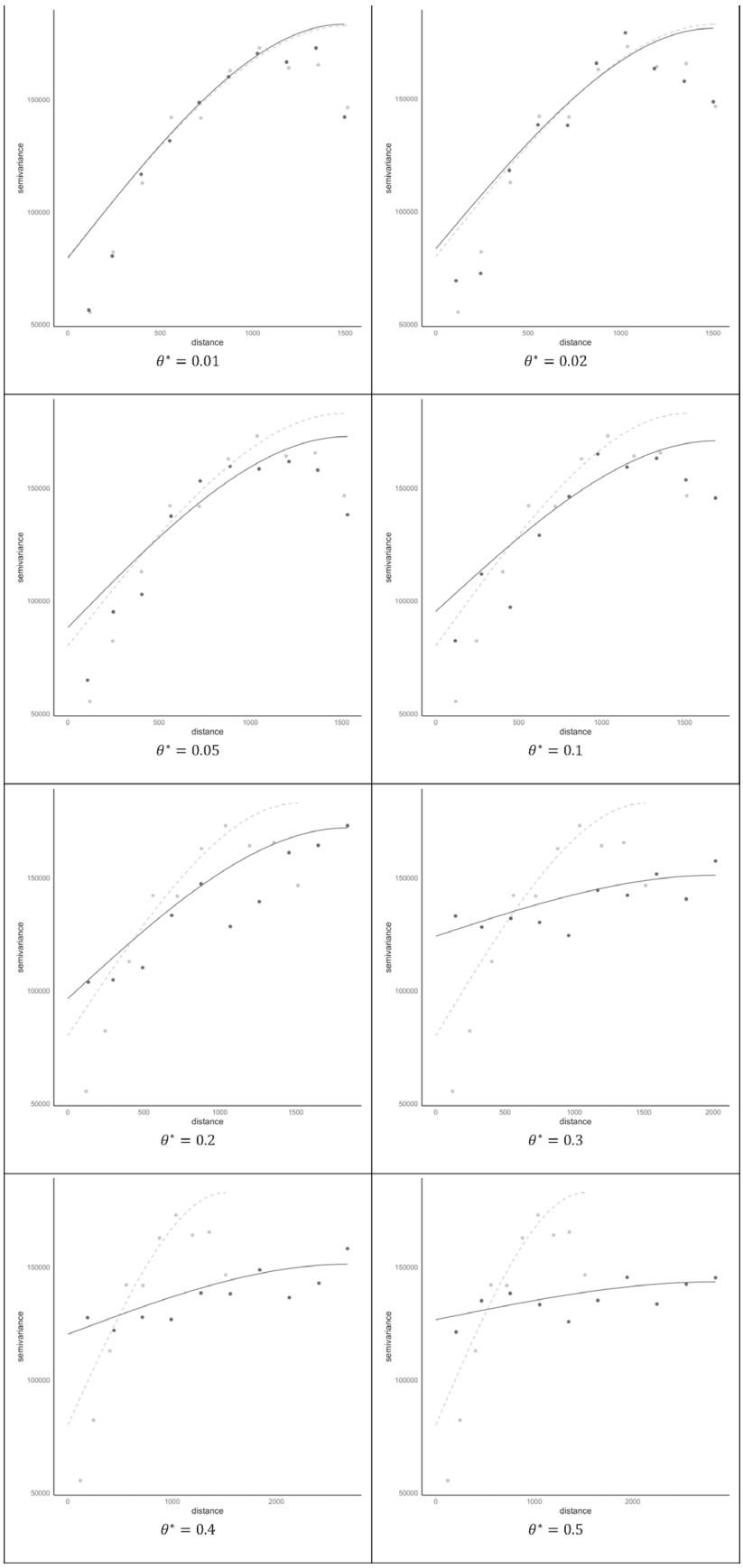

Figure 4. Effects of geo-masking on Gaussian semivariogram. True (black solid line) and estimated (gray dashed line) semivariograms after geo-masking at different levels of the displacement parameter θ^*. Distance is expressed in cm.

than an overestimation of the semivariogram. However, it is necessary to consider that Fig. 1 reports the expected value of the bias as it results from the theoretical expected value of the distance after contamination. In contrast, in the cases reported in Fig. 4 the distance used in the calculation is not an expected value, rather it derives from a single observation of the real data.

More in detail, Table 1 reports the effects of displacement on the various semivariogram parameters, as measured by the ratio between the parameters' estimation on true data set and the same on the geo-masked data set. In reading the table, consider that values greater than 1 imply an overestimation error due to geo-masking, while values less than 1 imply an underestimation.

Apart from an obvious trend for all parameters of showing an increasing bias when θ^* increases, the table clearly shows that the nugget parameter is always overestimated on the geo-masked data. In contrast, the partial sill is always underestimated (unless $\theta^* = 0.01$). Finally, the range is always overestimated when $\theta^* > 0.02$.

Fig. 5 shows visually the effects of geo-masking on the kriging predicted values when the parameter θ^* increases.

The visual inspection of the various graphs clearly shows that the estimation can tolerate low values of geo-masking displacement of about 1%. However, as soon as the displacement parameter is greater than 0.01, the image appears substantially different and when $\theta^* > 0.02$ it shares only a vague resemblance with the true one hiding important features of the map. The maps shown in Fig. 5 refer to a one random displacement in which the points are repositioned according to the geomasking logic described above. In order to show that our results are consistent, we simulated the geomasking procedure 500 times and we calculated the mean absolute error of the estimated model after geo-masking at different levels of θ^*. Fig. 6 shows that the Mean Absolute Error increases with the displacement parameter and it stabilizes after the value of $\theta^* = 0.3$.

In addition, in Fig. 7, we report the spatial distribution of the Mean Absolute Error after geo-masking. There is a phenomenon of underestimation in areas where the original zinc values are low and, conversely, of overestimation in areas where the values are high. This phenomenon is more visible as the displacement error increases, in line with the previous comments. Beyond the level of $\theta^* = 0.3$, the errors and their spatial distribution are very similar.

Table 1. Ratio between the estimation of the various semivariogram parameters on real data and the same estimated on the geo-masked data set for different levels of the displacement parameter θ^*

Displacement parameter θ^*	Nugget τ^2	Partial Sill $\sigma = 0.3$	Range α
0.01	0.98	0.99	1.00
0.02	0.98	1.00	1.00
0.05	0.88	1.03	0.95
0.1	0.78	1.08	0.91
0.2	0.63	1.22	0.80
0.3	0.57	1.29	0.68
0.4	0.66	1.25	0.60
0.5	0.65	1.26	0.51

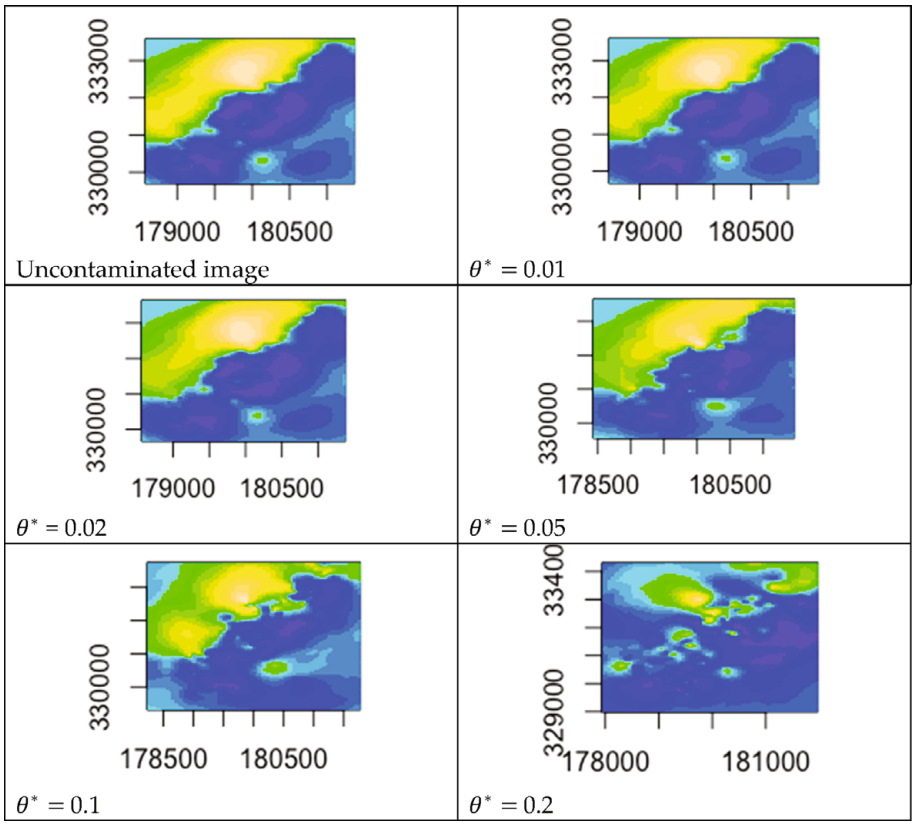

Figure 5. Effects of geo-masking on Gaussian semivariogram. True image of the predicted values and the after geo-masking at different levels of the displacement error θ*.

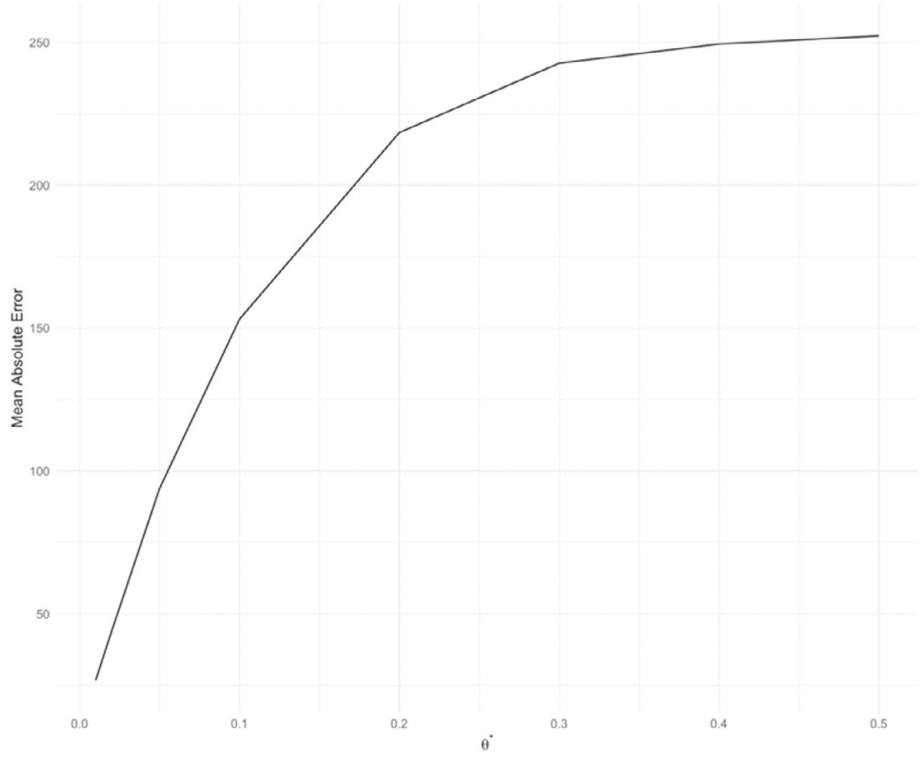

Figure 6. Mean absolute error of the estimated model after geo-masking at different levels of the displacement error θ*.

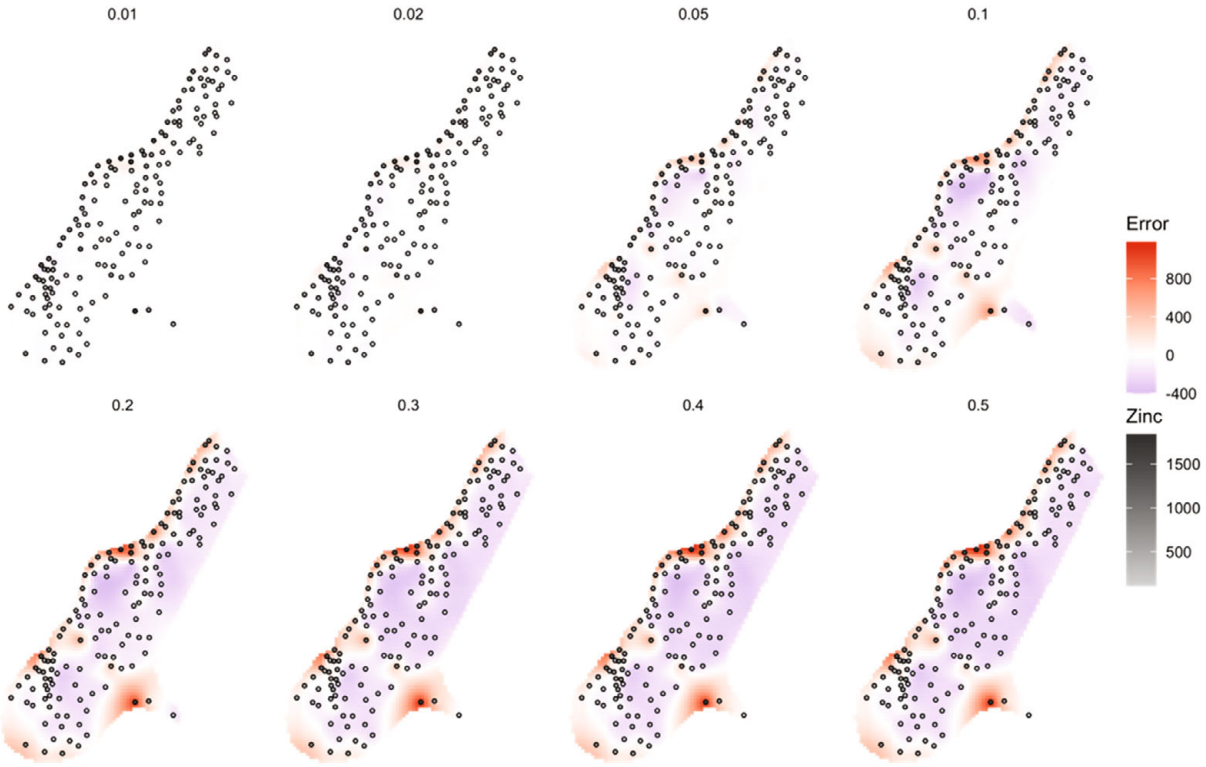

Figure 7. Maps of the true value of zinc (gray scale points) and mean absolute error of the estimated model after geo-masking at different levels of the displacement error θ^*. Red areas refer to positive errors while blue areas refer to negative errors.

Effects of geo-masking on the kriging variance

Let us, finally, turn our attention to consider the effects of geo-masking on the estimation variance (or *Kriging variance*). In the case of a simple kriging when the expected value of the underlying process is known and equal to 0 it is rather straightforward to show these effects.

In fact, in this case, the kriging variance can be expressed as:

$$\sigma_{sk}^2 = \sigma^2 - \sigma^T \Sigma^{-1} \sigma, \tag{7}$$

(Shabenberger and Gotway 2005, p. 224) with σ_{sk}^2 the simple kriging variance, $\sigma^T = \text{Cov}\left[Z(s_0)\, Z(s)\right]$ a vector of covariances between all sampled points and the unsampled point, say s_0. Without loss of generality, we are assuming that the point to be predicted is located at the origin of a unitary square. Finally, in Equation (7), $\Sigma = \text{Var}[Z(s)]$ represents the variance–covariance matrix between the sample points which, if we assume again a Gaussian covariance function, can be computed using Equation (1). In this case we can rather straightforwardly calculate the true σ_{sk}^2 in any experimental situation, while the kriging variance after geo-masking will be given by:

$$\overline{\sigma}_{sk}^2 = \sigma^2 - \overline{\sigma}^T \overline{\Sigma}^{-1} \overline{\sigma}, \tag{8}$$

where $\overline{\sigma}$ and $\overline{\Sigma}$ are now calculated with reference to the expected distances after geo-masking rather than to the true distances. In the case of the pairwise distances which are required for the calculation of $\overline{\Sigma}$ we can use again the result reported in the Appendix.

Furthermore, in this case, we have:

$$\sigma^T = Cov\left[Z(s_0)\, Z(s)\right] = \sigma^2 \exp\left\{-3\frac{d_0^2}{\alpha^2}\right\} + \tau^2, \tag{9}$$

where d_0^2 now represents the distance between each point and the unsampled point s_0. After geo-masking this expression becomes.

$$\overline{\sigma}^T = \sigma^2 \exp\left\{-3\frac{\overline{d}_0^2}{\alpha^2}\right\} + \tau^2. \tag{10}$$

From the Appendix we have that the expected value of $\overline{d}_0^2$ is:

$$E\left(\overline{d}_0^2\right) = d^2 + \frac{(\theta^*)^2}{3}, \tag{11}$$

and, substituting this value into Equation (10), we have:

$$\overline{\sigma}^T = \sigma^2 \exp\left\{-3\frac{d_0^2 + \frac{(\theta^*)^2}{3}}{\alpha^2}\right\} + \tau^2, \tag{12}$$

that can be easily computed.

The ratio between the kriging variance before and after geo-masking shows the efficiency loss in the prediction due to the geo-masking procedure that can be measured by the term:

$$EL = \frac{\sigma_{sk}^2}{\overline{\sigma}_{sk}^2}. \tag{13}$$

Fig. 8 reports the behavior of the efficiency loss for different levels of θ^* in various experimentally controlled situations. The true points are assumed again to be laid on a unitary square and to obey a Complete Spatial Randomness pattern (Diggle 2013). Points are generated by two independent uniform distributions from −0.5 to 0.5 in the two directions. Because of this, likewise the results reported in Section 2, the parameter θ^* ranges now between 0 and $\sqrt{2}$ in the unitary square. We considered again the semivariogram described in Equation (2) with the parameters' value set to $\alpha = 0.4$, $\sigma^2 = 0.18$, $\tau^2 = 0.09$.

The graph clearly shows how the efficiency decreases sharply already at low values of θ^*. In particular, the kriging variance drops down to 0.012% when $\theta^* = 0.1$ (7% of the maximum displacement distance), and it records a dramatic when θ^* goes down to 0.2.

In these conditions, apart from the bias observed in Sections 2 and 3, the semivariogram estimators becomes extremely inefficient when observed on even moderately geo-masked data points and so unreliable tools to perform spatial prediction and inference.

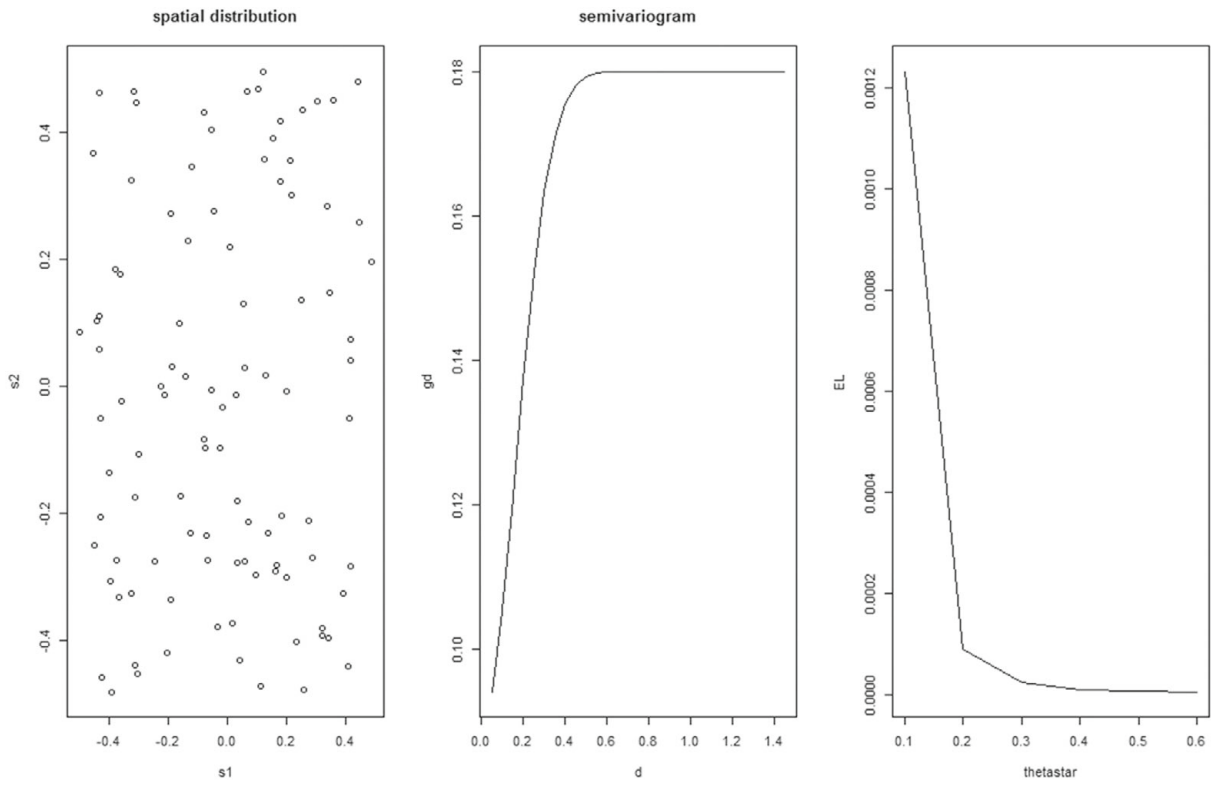

Figure 8. Complete spatial randomness spatial distribution of 100 points (left graph), associated Gaussian semivariogram (graph in the center) and efficiency loss due to geo-masking in the kriging variance for prediction of a point located in the origin (right graph) plotted against the displacement parameter θ^*.

Conclusions

Geostatistical methods are very common geographical tools employed in many disciplines and it is easy to forecast that they will increase their popularity with the diffusion of micro-data linked to the widespread use of alternative data sources (Arbia, 2021; Arbia et al., 2021).

However, their used are undermined by the fact that in many situations there is an uncertainty about the true position of the individual points due to the need to geo-mask the individual's position to preserve their confidentiality.

In this paper we have assumed that the geo-masking process only affects the parameters' values and not the shape of the semivariogram. Under this hypothesis we have shown that the semivariogram parameters estimators are biased and inefficient and prediction becomes unreliable. However, when the masking random mechanism is known, its parameters can be incorporated in the formal analysis and suggest what are the limits within which geo-masking can be tolerated by the geostatistical methods and when they are dramatically distorted. If we adopt the common random-angle random-distance geo-masking method, we have shown that the crucial geo-masking parameter is the maximum displacement distance while the random angle of displacement produces no distortions. In the case of a simple kriging, based on theoretical results and the exam of real and artificial Monte Carlo-generated data, our analysis shows that when this crucial parameter overcomes the value of 0.1 (that is about 7% of the maximum displacement distance) the consequences are dramatic both on the estimation of the semivariogram and of its parameters and on the estimation variance (or *Kriging variance*). The results of this paper should raise researchers' awareness of the possible devastating consequences of the presence

of geo-masking in running geostatistical analyses. Furthermore, they could be used by data producers in order to calibrate the optimal value of θ^* and to communicate it to the practitioners and researchers so that they could anticipate the expected level of accuracy of their analyses.

While the theoretical analysis presented in this paper is limited to simple kriging, to Gaussian semivariogram and to a specific geo-masking mechanism, the results obtained are rather general and the framework adopted here could be extended to consider universal kriging, different semivariogram formulations and different geo-masking mechanism if required.

Appendix: Proof of the effects of the random-direction random-distance geo-masking on inter-point distances

In order to examine the effects of geo-masking on the calculation of distances in two dimensions, let us consider two generic points of coordinates (x_i,y_i) and (x_j,y_j). The true pairwise Euclidean distance on the true positions are defined as:

$$d_{i,j}^2 = \left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2. \tag{A1}$$

In contrast, the distance between two points after a geo-masking with the *random direction random distance* method can be expressed, is defined as:

$$\overline{d}_{i,j}^{\ 2} = \left(x_i + \theta_{i,1}\cos\delta_{i,1} - x_j - \theta_{j,1}\cos\delta_{j,1}\right)^2 + \left(y_i + \theta_{i,2}\sin\delta_{i,2} - y_j - \theta_{j,2}\sin\delta_{j,2}\right)^2, \tag{A2}$$

by using the polar coordinates. In Equation (A2) $\theta_{i,.}$ and $\theta_{j,.}$ are independent realizations of the random variable θ_i and similarly and $\delta_{i,.}$ and $\delta_{j,.}$ are independent realizations of the random variable δ_i so that $\theta_{i,.} \approx U\left(0,\theta^*\right)$ and $\delta_{i,.} \approx U\left(0,360°\right)$, having defined θ^* as the maximum distance error and $\theta_{i,.}$ and $\delta_{i,.}$ independent of one another and independent of the variables observed. Squaring the two terms in the RHS, from Equation (A2) we have:

$$\begin{aligned}\overline{d}_{i,j}^{\ 2} = {} & x_i^2 + \theta_{i,1}^2\left(\cos\delta_{i,1}\right)^2 + x_j^2 + \theta_{j,1}^2\left(\cos\delta_{j,1}\right)^2 + 2x_i\theta_{i,1}\cos\delta_{i,1} - 2x_ix_j - 2x_i\theta_{j,1}\cos\delta_{\mathrm{j},1} \\ & - 2x_j\theta_{i,1}\cos\delta_{i,1} - 2\theta_{i,1}\theta_{j,1}\cos\delta_{i,1}\cos\delta_{j,1} + 2x_j\theta_{j,1}\cos\delta_{j,1} + y_i^2 + \theta_{i,2}^2\left(\sin\delta_{i,2}\right)^2 + y_j^2 \\ & + \theta_{j,2}^2\left(\sin\delta_{j,2}\right)^2 + 2y_i\theta_{i,2}\sin\delta_{i,2} - 2y_iy_j - 2y_i\theta_{j,2}\sin\delta_{j,2} - 2y_j\theta_{i,2}\sin\delta_{i,2} \\ & - 2\theta_{i,2}\theta_{j,2}\sin\delta_{i,2}\sin\delta_{j,2} + 2y_j\theta_{j,2}\sin\delta_{j,2}.\end{aligned} \tag{A3}$$

Hence:

$$\begin{aligned}\overline{d}_{i,j}^{\ 2} = {} & d_{i,j}^2 + \theta_{i,1}^2\left(\cos\delta_{i,1}\right)^2 + \theta_{j,1}^2\left(\cos\delta_{j,1}\right)^2 + 2x_i\theta_{i,1}\cos\delta_{i,1} - 2x_i\theta_{j,1}\cos\delta_{j,1} - 2x_j\theta_{i,1}\cos\delta_{i,1} \\ & - 2\theta_{i,1}\theta_{j,1}\cos\delta_{i,1}\cos\delta_{j,1} + 2x_j\theta_{j,1}\cos\delta_{j,1} + \theta_{i,2}^2\left(\sin\delta_{i,2}\right)^2 + \theta_{j,2}^2\left(\sin\delta_{j,2}\right)^2 \\ & + 2y_i\theta_{i,2}\sin\delta_{i,2} - 2y_i\theta_{j,2}\sin\delta_{j,2} - 2y_j\theta_{i,2}\sin\delta_{i,2} - 2\theta_{i,2}\theta_{j,2}\sin\delta_{i,2}\sin\delta_{j,2} + 2y_j\theta_{j,2}\sin\delta_{j,2}.\end{aligned} \tag{A4}$$

Considering that $E\left(\cos\delta_{i,.}\right) = E\left(\sin\delta_{i,.}\right) = 0$ and the hypothesis of independence of $\delta_{i,.}$ and $\theta_{i,.}$, the expectation of Equation (A4) is:

$$\begin{aligned}E\left(\overline{d}_{i,j}^{\ 2}\right) = {} & d_{i,j}^2 + E\left(\theta_{i,1}^2\right)E\left[\left(\cos\delta_{i,1}\right)^2\right]E + E\left(\theta_{i,2}^2\right)E\left[\left(\sin\delta_{i,2}\right)^2\right] \\ & + E\left(\theta_{j,1}^2\right)E\left[\left(\cos\delta_{j,1}\right)^2\right]E + E\left(\theta_{j,2}^2\right)E\left[\left(\sin\delta_{j,2}\right)^2\right].\end{aligned} \tag{A5}$$

Since $E\left[\left(\cos\delta_{i,.}\right)^2\right] = E\left[\left(\sin\delta_{i,.}\right)^2\right] = 1/2$ and $E\left(\theta_{i,.}^2\right) = (\theta^*)^2/3$ we have

$$E\left(\bar{d}_{i,j}{}^2\right) = d_{i,j}^2 + \frac{2}{3}\left(\theta^*\right)^2. \tag{A6}$$

Let us now turn to consider the effects of a *random-distance random-direction* geo-masking on the calculation of distances between each observed point and an unobserved point where we wish to predict the value of some variables using a kriging procedure. Without loss of generality let us assume that the point to be predicted is located at the origin of unitary square, say $s_0 = (0,0)$. Under this hypothesis, the true squared distance between a generic point of (x_i, y_i) and the unobserved point before geo-masking can be calculated as $d_{ij}^2 = \left(x_i{}^2 + y_i^2\right)$, while, after geo-masking, the observed distance becomes:

$$\bar{d}_{ij}^2 = \left(x_i + \theta_{i,1}\cos\delta_{i,1}\right)^2 + \left(x_i + \theta_{i,2}\cos\delta_{i,2}\right)^2. \tag{A7}$$

From Equation (A7) we can express the geo-masking expected value as:

$$\bar{d}_{i,j}{}^2 = d_{i,j}^2 + \theta_{i,1}^2\left(\cos\delta_{i,1}\right)^2 + \theta_{i,2}^2\left(\sin\delta_{i,2}\right)^2 + 2x_i\theta_{i,1}\cos\delta_{i,1} + 2y_i\theta_{i,2}\sin\delta_{i,2}. \tag{A8}$$

If we take the expected value of the LHS of Equation (A8) we have as in (A5):

$$E\left(\bar{d}_{i,j}{}^2\right) = d_{i,j}^2 + \frac{2}{3}\left(\theta^*\right)^2.$$

We remind that:

$$\begin{aligned} E\left(\cos\delta_{i,.}\right) &= \int_0^{360}\cos\left(\delta_{i,.}\right) f\left(\delta_{i,.}\right) d\delta = \frac{1}{360}\int_0^{360}\cos\left(\delta_{i,.}\right) d\delta = \frac{1}{360}\left[\sin\delta_{i,.}\right]_0^{360} \\ &= 0 = E\left(\sin\delta_{i,.}\right). \end{aligned} \tag{A9}$$

Finally, from the hypothesis of uniform distribution of $\theta_{i,.}$, we can express its expected value and variance as $E\left(\theta_{i,.}\right) = \frac{\theta^*}{2}$ and $\mathrm{Var}\left(\theta_{i,.}\right) = \frac{\left(\theta^*\right)^2}{12}$, respectively. Hence

$$E\left(\theta_{i,.}{}^2\right) = \mathrm{Var}\left(\theta_{i,.}\right) + E\left(\theta_{i,.}\right)^2 = \frac{(\theta^*)^2}{12} + \frac{(\theta^*)^2}{4} = \frac{(\theta^*)^2}{3}, \tag{A10}$$

which produces the result exploited in Equation (A6).

REFERENCES

Allhouse, W. B., M. K. Fitch, K. H. Hapton, D. C. Gesnik, I. A. Doherty, P. A. Leone, M. L. Serre, and W. C. Miller. (2010). "Geo-Masking Sensitive Health Data and Privacy Protection: An Evaluation Using an E911 Database." *Geocarto International* 25(6), 443–52.

Arbia, G. (2021). *Statistics, Society and New Empiricism in the Era of Big Data*. Springerbrief in Statistics. Cham: Springer Verlag.

Arbia, G., G. Espa, and D. Giuliani. (2016). "Dirty Spatial Econometrics." *Annals of Regional Science* 56, 177–89.

Arbia, G., G. Espa, and D. Giuliani. (2021). *Spatial Microeconometrics*. London: Routledge.

Banerjee, S., B. P. Carlin, and A. E. Gelfand. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. London: Chapman & Hall/CRC.

Burgert, C. R., J. Colston, T. Roy, and B. Zachary. (2013). "Geographic Displacement Procedure and Georeferenced Data Release Policy for the Demographic and Health Surveys." DHS Spatial Analysis Report No. 7.

Burrough, P. A., and R. A. McDonnell. (1998). *Principles of Geographical Information Systems*, 2nd ed. Oxford: Oxford University Press.

Cassa, C. A., S. J. Grannis, J. M. Overhage, and K. D. Mandl. (2006). "A Context-Sensitive Approach to Anonymizing Spatial Surveillance Data: Impact on Outbreak Detection." *Journal the of American Medical Informatics Association* 13(2), 160–5.

CEBM. (2020). "Sars-Cov-2 Viral Load and the Severity of COVID-19." Center for Evidence-Based Medicine, 26th March, 2020.

Collins, B. (2011). *Boundary Respecting Point Displacement, Python Script*. Arlington, VA: Blue Raster, LLC.

Cressie, N., and J. Kornak. (2003). "Spatial Statistics in the Presence of Location Error with an Application to Remote Sensing of the Environment." *Statistical Science* 18(4), 436–56.

Diggle, P. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*, 3rd ed. Boca Raton: CRC Press.

Diggle, P., and E. Giorgi. (2021). *Model-Based Geostatistics for Global Public Health Methods and Applications*. London: Chapman and Hall/CRC.

Diggle, P., and P. J. Ribeiro. (2007). *Model-Based Geostatistics*. New York, NY: Springer Verlag.

Fronterrè, C., E. Giorgi, and P. Diggle. (2018). "Geostatistical Inference in the Presence of Geo-Masking: A Composite-Likelihood Approach." *Spatial Statistics* 28, 319–330.

Gabrosek, J., and N. Cressie. (2002). "The Effect on Attribute Prediction of Location Uncertainty in Spatial Data." *Geographical Analysis* 34(3), 262–285.

Gao, S., J. Rao, X. Liu, Y. Kang, Q. Huang, and J. App. (2019). "Exploring the Effectiveness of Geo-Masking Techniques for Protecting the Geoprivacy of Twitter Users." *Journal of Spatial Information Science* 19, 105–29.

Gething, P. W., A. M. Noor, P. W. Gikandi, S. I. Hay, M. S. Nixon, R. W. Snow, and P. Goovaerts. (2005). "Geostatistical Analysis of Disease Data: Estimation of Cancer Mortality Risk from Empirical Frequencies Using Poisson Kriging." *International Journal of Health Geography* 4, 31.

Goovaerts, P. (2008). "Geostatistical Analysis of Health Data: State-of-the-Art and Perspectives." In *geoENV VI – Geostatistics for Environmental Applications*. Quantitative Geology and Geostatistics Vol 15, edited by A. Soares, M. J. Pereira, and R. Dimitrakopoulos. Dordrecht: Springer.

Grosh, E. and J. M. Munoz. (1996). A Manual for Planning and Implementing the Living Standards Measurement Study Survey. Technical Report No. LSM126, The World Bank.

Hampton, K. H., M. K. Fitch, W. B. Allshouse, I. A. Doherty, D. C. Gesink, P. A. Leone, M. L. Serre, and W. C. Miller. (2010). "Mapping Health Data: Improved Privacy Protection with Donut Method Geo-Masking." *American Journal of Epidemiology* 172(9), 1062–9.

IFNC. (2015). http://www.sian.it/inventarioforestale/jsp/home_en.jsp

Kerry, R., P. Goovaerts, R. P. Haining, and V. Ceccato. (2010). "Applying Geostatistical Analysis to Crime Data: Car-Related Thefts in the Baltic States." *Geographical Analysis* 42(1), 53–77.

Kyriakidis, P. C. (2010). "A Geostatistical Framework for Area-to-Point Spatial Interpolation." *Geographical Analysis* 36(3), 259–89.

Kyriakidis, P. C., and E.-H. Yoo. (2005). "Geostatistical Prediction and Simulation of Point Values from Areal Data." *Geographical Analysis* 37(2), 124–51.

Mcroberts, R. E., G. R. Holden, M. Nelson, and D. D. Gormanson. (2005). "Using Satellite Imagery as Ancillary Data for Increasing the Precision of Estimates for the Forest Inventory and Analysis Program of the USDA Forest Service." *Canadian Journal of Forest Research* 35, 12.

Montero, J.-M., G. Fernàndez-Avilés, and J. Mateu. (2015). *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging*. London: Wiley.

Oliver, M. A., R. Webster, C. Lajaunie, K. R. Muir, S. E. Parkes, A. H. Cameron, M. C. G. Stevens, and J. R. Mann. (1998). "Binomial Cokriging for Estimating and Mapping the Risk of Childhood Cancer." *IMA Journal of Mathematics Applied in Medicine and Biology* 15, 279–97.

Pawitan, G., and D. G. Steel. (2006). "Exploring a Relationship between Aggregate and Individual Levels Spatial Data through Semivariogram Models." *Geographical Analysis* 38(3), 310–25.

Santos, A., N. Medeiros, G. Santos, and J. Lisboa. (2017). "Use of Geostatistics on Absolute Positional Accuracy Assessment of Geospatial Data." *Boletim de Ciências Geodésicas* 23(3), 405–18.

Seidl, D. E., P. Jankowski, and K. Clarke. (2018). "Privacy and False Identification Risk in Geo-Masking Techniques." *Geographical Analysis* 50(3), 280–97.

Shabenberger, O. and C. A. Gotway. (2005). *Statistical Methods for Spatial Data Analysis*. London: Chapmann & Hall/CRC.

Singanayagam, A., S. Hakki, J. Dunning, K. J. Madon, M. A. Crone, A. Koycheva, N. Derqui-Fernandez, et al. (2022). "Community Transmission and Viral Load Kinetics of the SARS-CoV-2 Delta (B.1.617.2) Variant in Vaccinated and Unvaccinated Individuals in the UK: A Prospective, Longitudinal, Cohort Study." *The Lancet Infectious diseases* 22, 183–95. https://doi.org/10.1016/S1473-3099(21)00648-4

Webster, R., M. A. Oliver, K. R. Muir, and J. R. Mann. (1994). "Kriging the Local Risk of a Rare Disease from a Register of Diagnoses." *Geographical Analysis* 26(2), 168–85.

Zandbergen, P. A. (2014). "Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data." *Advances in Medicine* 2014, 567049.

Zawadzki, J., C. J. Cieszewski, M. Zasada, and R. C. Lowe. (2005). "Applying Geostatistics for Investigations of Forest Ecosystems Using Remote Sensing Imagery." *Silva Fennica* 39(4), 599–617.

Zhang, J., and P. K. Roger. (2000). "A Geostatistical Approach to Modelling Positional Errors in Vector Data." *Transactions in GIS* 4(2000), 145–59.

Zhang, S., S. Freundshuh, K. Lenzer, and P. A. Zandbergen. (2015). "The Location Swapping Method for Geo-Masking." *Cartography and Geographic Information Science* 44, 22–34.

Chapter 4

Modelling

Robust Measures of Spatial Correlation

Robust Measures of Spatial Correlation

Vincenzo Nardelli, Giuseppe Arbia

Abstract

Statistical measures, across various disciplines, are vulnerable to the effects of outliers. Spatial correlation coefficients, critical in the assessment of spatial data, remain susceptible to this inherent flaw. In contexts where data is sourced from diverse domains—ranging from regular lattices, like satellite imagery, to non-lattice constructs such as administrative divisions—it's not uncommon to witness a few anomalous data points. Such outliers can skew the broader analytical landscape, often masking significant spatial attributes. This paper embarks on a mission to enhance the resilience of traditional spatial correlation metrics, specifically the Moran coefficient (MC), Geary's contiguity ratio (CR), and the approximate profile likelihood estimator (APLE). Drawing inspiration from established analytical paradigms, our research harnesses the power of influence function studies to examine the robustness of traditional methods against novel alternatives. Employing Monte Carlo simulations, we simulated outlier scenarios into spatial data sets to test the mettle of these metrics.

Keywords: Spatial correlation; Influence functions; Robust estimation.

1 Introduction

Most statistical measures are very sensitive to outliers and spatial correlation coefficients are not an exception to this rule. Indeed, spatial outliers are very common in practice when data are observed both on regular lattices (e. g. in satellite images) and in non-lattice data such as administrative partitions like countries or regions. In these cases, the presence of few exceptional observations may dramatically distort the picture and hide interesting spatial features. In this paper we introduce methods for robustizing traditional spatial correlation measures, such as the Moran coefficient (MC; see Moran (1948)), Geary's contiguity ratio (CR; see Geary (1954)) and the approximate profile likelihood estimator (APLE; see Li et al. (2007, 2012)). Although not in the same way, all three measures are sensitive to observations that may disproportionately influence the measurement of spatial dependence. Following the traditional approach, we will base our analysis on the exam of influence function (Hampel, 1974) through which we will compare the robustness performances of the traditional measures with those of our proposed alternatives. For the sake of comparisons, we will make use of Monte Carlo experiments with which we simulate sets of spatial

data, where we surreptitiously introduce different simulated outlier conditions. As it is well known, robust estimation is intrinsically connected with outlier detection. Therefore, after introducing our alternative measures, we will use them to develop procedures for detecting spatial outliers. The following Section 2 will concern the presentation of some spatial correlation measures. Sections 3 and 4 will be devoted to influence functions and to robust estimation respectively. Section 5 will contain the conclusions.

2 Some measures of spatial correlation

At the heart of all measures of spatial correlation studied in the literature we find the definition of the so-called weight matrix (W) which accomplishes the task of describing the topology of the spatial system on which the data are laid. As we will see later, it also plays a fundamental role in the analysis of the effect of outliers on spatial correlation measures. Suppose we have n observations of a random variable Z say $Z = (z_1, z_2, \ldots, z_n)$, which, without loss of generality, are assumed centered around the mean and distributed on (possibly irregoular) lattice locations. The generic entry $w_{ij} \in W$, expresses the level of connectedness

between location i and location j, where

$$w_{ij} = \begin{cases} 0 & \text{if } i = j \\ > 0 & \text{if } j \in N(i) \\ 0 & \text{otherwise} \end{cases}$$

with $N(i)$ the set of locations connected with location i.

Consequently, $\sum_{j=1}^{n} w_{ij} = \eta_i$ is the connectivity of location i (or weighted outdegree in the graph theory terminology. See Bang-Jensen and Gutin (2007)), and $\bar{\eta} = n^{-1} \sum_{i=1}^{n} \eta_i$ is the average connectivity of the spatial system. W is often row-standardized so that $\eta_i = 1$ for each i and $\bar{\eta} = n$. In the reminder of the paper, the symbol W will indicate the row-standardized version. Given these definitions the weighted average of the neighbours of location i :

$$L[z_i] = \sum_{i=1}^{n} w_{ij} z_j \tag{1}$$

assumes the role of the spatially lagged variable by analogy to the time series definition. Generalizing, we have $L[Z] = WZ$.

The Moran coefficient (Moran, 1950) can then be defined as:

$$MC = \frac{\sum_{i=1}^{n} (z_i) L[z_i]}{\sum_{i=1}^{n} (z_i)^2} = \frac{Z^T L(Z)}{Z^T Z} \tag{2}$$

Equation 2 is the ratio between the spatial autocovariance and the variance of X. However, it is improperly referred to as a spatial correlation coefficient. Indeed, it assumes the form of a correlation only if $\mathrm{Var}(x) = \mathrm{Var}(L[x])$, which is

not the case unless in trivial situations. As a consequence, its range is narrower than the interval $[-1;1]$ (see Arbia et al. (1989)) and it depends on the extreme eigenvalues of W (Griffith, 2010).

The APLE statistics (Calder and Cressie, 2007) was introduced to tackle one important limitation of MC: it is good estimator of the parameter of a spatial autoregressive model (Cressie, 1993) only in trivial cases of no practical interest. As an alternative, APLE assumes the following form:

$$APLE = \frac{1}{2}\frac{\left[Z^T W^T Z + Z^T W Z\right]}{Z^T W^T W Z + \mathrm{tr}\left(W^2\right) Z^T Z/n} = \frac{1}{2}\frac{\left[L(Z)^T Z + Z^T L(Z)\right]}{L(Z)^T L(Z) + \mathrm{tr}\left(W^2\right) Z^T Z/n} \tag{3}$$

which, when W is symmetric, boils down to a spatial autocovariance with a different normalizing factor in the dominator.

Finally, Geary's coefficient (Geary, 1954) is not a correlation measure, and is expressed as the ratio of two sums of squares:

$$GC = \frac{(2n\bar{\eta})^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(z_i - z_j\right)^2}{(n-1)^{-1}\sum_{i=1}^{n}(z)^2} \tag{4}$$

and, rather counterintuitively, falls in the range $[0;2]$ revealing negative spatial correlation if greater than 1, positive if lower than 1 , and no correlation if equal to 1.

Although different if used in a descriptive context, MC and GC are inferentially equivalent.

3 Influence functions in space

In general, if we define $\hat{\theta}$ as an estimator of a generic parameter θ, based on n observations, and $\hat{\theta}_+$an estimator of the same form of $\hat{\theta}$ based on the same n observations, but also on an additional observation x_o, the finite sample version of Hampel's influence function (Hampel, 1974) can be defined as $I_+\left(\theta, x_o\right) = (n+1)\left(\hat{\theta}_+ - \hat{\theta}\right)$. This quantity, in general, depends only on the amount of contamination x_o. However, when considering data distributed in space, the influence function depends also on the location where the contamination is observed and on its connection with the neighboring locations. Intuitively, given the nature of dependence between spatial data, if the contaminated location is strongly connected with other locations (that is, it is a dominant unit according to the definition of (Pesaran and Yang, 2020)), the influence of x_o will be stronger than in the case of loosely connected units. Indeed, in this case its effect propagates also to the neighboring units and hence it corrupts more substantially the spatial correlation parameters.

In the case of MC and APLE coefficients, some theoretical results could be derived to support such an intuition. At the basis of both MC and APLE statistics, indeed, is the calculation of the spatial autocovariance appearing in their

numerator, say $\gamma = n^{-1} Z^T W Z$. Let us now consider an additional observation z_0, the augmented vector of observations $\overline{\boldsymbol{z}}^T \equiv [z_1, z_2, \dots z_n, z_0]$ and the term $\omega_{0i} \in \omega_0$ defined as the generic element of an n-by- 1 column vector summarizing the connectivity of the additional observation with the existing ones. Let us also express $\varphi_0 = \sum_{i=1}^{n} \omega_{0i}$ as the sum of the connections of the additional unit. We can then express the weight matrix, including the additional information, as:

$$\bar{W} \equiv \begin{bmatrix} \boldsymbol{W} & \boldsymbol{\omega}_0 \\ \boldsymbol{\omega}_0^T & 0 \end{bmatrix} \tag{5}$$

After some straightforward algebra, the spatial autocovariance estimated with an additional unit can be expressed as:

$$\hat{\gamma}_+ = n^{-1} \bar{Z}^T \overline{\boldsymbol{W}} \bar{Z} = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} z_i z_j + n^{-1} 2 z_0 \sum_{i=1}^{n} \omega_{0i} z_i \tag{6}$$

The I_+empirical influence function of γ can then be written as:

$$I_+ (z_0; \gamma) = (n+1) (\hat{\gamma}_+ - \hat{\gamma}) = \frac{(n+1)}{n} 2 z_0 \sum_{i=1}^{n} \omega_{0i} z_i \tag{7}$$

which is an increasing function not only of the perturbating observation z_0, but also of its connections with all the other units (ω_{0i}) and of the values observed in the connected units.

In Figure 1, the simulated influence functions for MC, GC, and APLE are presented. Notably, MC and GC overlap precisely, indicating they possess identical influence functions, while APLE demonstrates reduced robustness to outliers.

4 Robust estimation of spatial correlation

4.1 Specification of robust estimators

The unboundedness of the influence function of the three spatial correlation measures discussed in the previous section, leads us to suggests various robust alternatives to protect against the possible distortions due to the presence of outliers.

Preliminarily to the presentation of the various estimators, let us first introduce the notion of the robust spatial lag (RL), defined as the weighted median of the neighbours of x_i according to the topology described by the matrix W.

$$RL (x_i) = \text{Med} (x_j) ; j \in N(i) \tag{8}$$

We then consider the following alternative estimators:

a) MC using the robust spatial lag definition in place of the spatial lag definition:

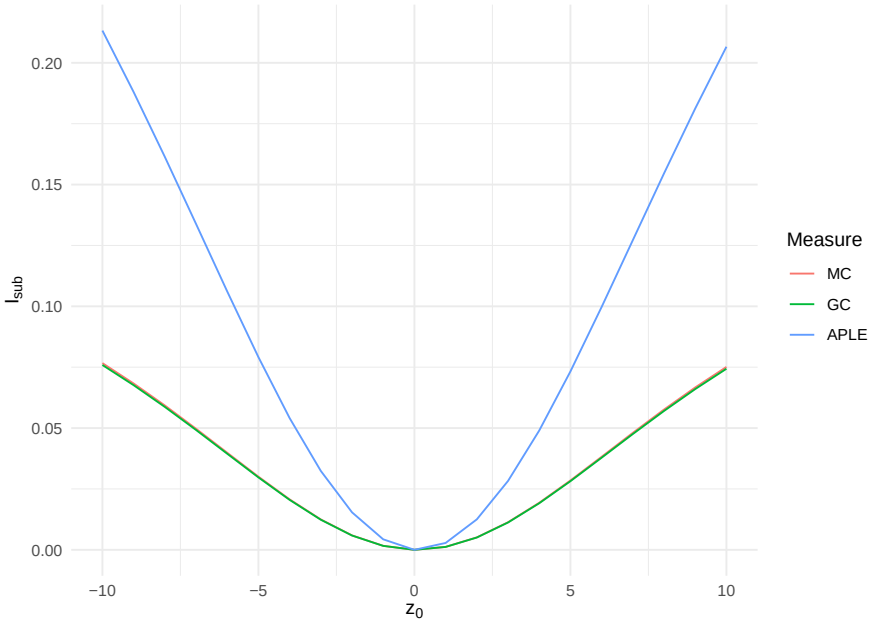

Figure 1: Influence function for Moran Coefficient (MC), Geary's Coefficient (GC), and APLE.

$$RMC = \frac{\sum_{i=1}^{n} (z_i) (RL[z_i]))}{\sum_{i=1}^{n} (z_i)^2} \tag{9}$$

b) APLE using the robust spatial lag definition in place of the spatial lag definition:

$$RAPLE = \frac{1}{2} \frac{\left[RL(Z)^T Z + Z^T RL(Z)\right]}{RL(Z)^T RL(Z) + \mathrm{tr}\,(W^2)\, Z^T Z/n} \tag{10}$$

c) GC using robust versions of the two sums of squares appearing in the numerator and respectively in the denominator:

$$RGC = \frac{(2n\bar{\eta})^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} |z_i - z_j|}{(n-1)^{-1} \sum_{i=1}^{n} |x_i|} \tag{11}$$

d) To introduce a further measure, let us recall the general notion of robust correlation proposed by Gnanadesikan and Kettenring (1972):

$$\frac{S(aX + bY)^2 - S(aX - bY)^2}{S(aX + bY)^2 + S(aX - bY)^2} \tag{12}$$

with $a = S(X)^{-1}, b = S(Y)^{-1}$, and S any robust measure of scale. If X is substituted by Z, Y by the spatially lagged value of Z and we opt the Median Absolute Deviation from the median (MAD) as a robust measure of scale, we obtain a further alternative as:

$$GK = \frac{MAD(aZ + L(Z))^2 - MAD(aZ - bL(Z))^2}{MAD(aZ + bL(Z))^2 + MAD(aZ - bL(Z))^2} \tag{13}$$

e) Finally, the last measure can be further robustized by using the robust spatial lag definition:

$$GK2 = \frac{MAD(aZ + LR(Z))^2 - MAD(aZ - bRL(Z))^2}{MAD(aZ + bRL(Z))^2 + MAD(aZ - bRL(Z))^2} \tag{14}$$

4.2 A Monte Carlo study of the estimators

The finite sample properties of all the suggested robust measures of spatial correlation will be now investigated. First, we derive the simulated empirical influence functions of the various measures. Secondly, we investigate the effects of using the various alternatives in different experimental situations.

4.2.1 Simulated empirical influence function of robust measures

In Figure 2, we examine the influence functions (IF) for various robust measures. When contrasted with the original measures presented in Figure 1, all the measures demonstrate enhanced robustness. Specifically, while RAPLE is notably the most sensitive to outliers, both RMC and RGC exhibit comparable resilience to extreme values. Amongst all, our proposed measures, GK and GK2, stand out; however, GK2 might be overly robust, potentially indicating a diminished sensitivity.

4.2.2 Performances of the robust estimators

In the light of the previous results, the performance of the various proposed estimators will be now studied through a simulation based on two different sample: a 10-by-10 and a 20-by20 regoular square lattice grid ($n = 100$ and $n = 400$). The data are originally generated with the following spatial autoregressive model (Cressie, 1993):

$$Z = \rho WZ + \epsilon$$

$$\epsilon \approx MVN\left(0, \sigma^2 I\right)$$

We considered three values of $\rho(-0.5; 0, 05)$ and, following Devlin et al. (1975), four different distributions for ϵ, namely:

a) Normal distribution

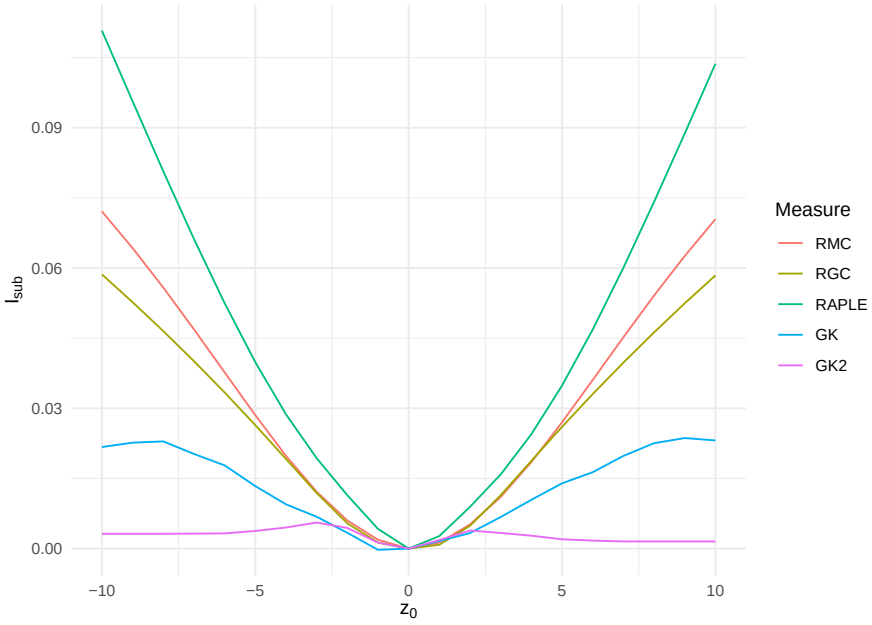

Figure 2: Influence function for Robust Moran Coefficient (RMC), Robust Geary's Coefficient (RGC), Robust APLE (RAPLE), GK and GK2 Coefficients.

b) Cauchy distribution

c) Laplace distribution

d) a mixture of two normal distributions $\left[0.9N\left(0, \sigma^2 I\right) + 0.1N\left(0, 9\sigma^2 I\right)\right]$

It's essential to highlight that while the first three distributions are symmetric and centered, the fourth one—the mixture of two normal distributions—is intrinsically asymmetric. This analysis represents a novel approach as it diverges from traditional testing methods. In real-world data scenarios, it is a common observation that data doesn't always follow normal distribution patterns. Furthermore, the presence of skewness—a lack of symmetry in data distribution—becomes especially pronounced in real-world datasets. By incorporating both centered and asymmetric distributions in our test, we present a more comprehensive and realistic examination, recognizing the complexities present in actual data.

Furthermore, to explore the effects of different connectivity in the spatial system, we also considered two definitions of the W matrix using the rook's case (where $\eta_i = 4$ for each i apart from those located at the edge of the lattice grid) and the queen's case (where $\eta_i = 8$ for each i apart from those located at the edge of the lattice grid) (Cressie, 1993). We repeated each experimental situation 1,000 times.

The performance of each estimator, as evaluated through Monte Carlo simulations, is displayed in Table 1.

Table 1: Percentage of simulation cases when the null of no spatial autocorrelation is rejected with $\rho = 0$, $n = 10$ and Queen W matrix.

Measure	Normal	Cauchy	Laplace	Mixture
MC	0.06	0.07	0.05	0.54
GC	0.05	0.06	0.05	0.77
APLE	0.06	0.06	0.05	0.55
RMC	0.06	0.06	0.05	0.34
RGC	0.05	0.06	0.05	0.58
RAPLE	0.06	0.05	0.04	0.35
GK	0.05	0.05	0.05	0.09
GK2	0.05	0.05	0.05	0.07

The appendix provides a comprehensive list of all simulation combinations. Observing the results for the Normal, Cauchy, and Laplace distributions, it is evident that all measures showcase robustness to outliers, consistently hovering around the value of 0.05. However, when analyzing the Mixture distribution, a distinct pattern emerges. Every measure, without exception, departs from the target 0.05 value. This indicates a significant susceptibility to the asymmetry and inherent complexities of data generated with the mixture distribution. Notably, while most measures show a pronounced departure, the proposed measures, GK and GK2, display remarkable resilience. Their superior performance over the other measures underscores the potential and efficacy of our proposed approach. This showcases not only the robustness of our proposed measures but also their superiority in terms of adaptability to diverse data distributions, particularly those that are skewed or mixed.

In Figure 3, we present the test power for all distributions, comparing the Weight Matrices. Initial observations from the results indicate that for the Normal, Cauchy, and Laplace distributions, all measures consistently exhibit robustness to outliers, irrespective of variations in either W or n. However, for the Mixture distribution, GK and GK2 stand out as significantly more robust.

The enhanced power of the mixture in the rook, when compared to the queen, is attributed to the variations in connectivity. As the number of neighbors increases, as seen in the rook scenario, the test's power diminishes. This observation aligns with previous findings demonstrated in the empirical influence function.

In Figure 4, we systematically showcase the test power for each distribution by comparing the effects of size. Notably, the patterns and observations derived from these results are strikingly similar to those presented in Figure 3, reinforcing the consistency and reliability of our findings across different metrics and evaluations. Within these findings, both GK and GK2 continue to demonstrate heightened robustness. To conclude, when considering the influence functions,

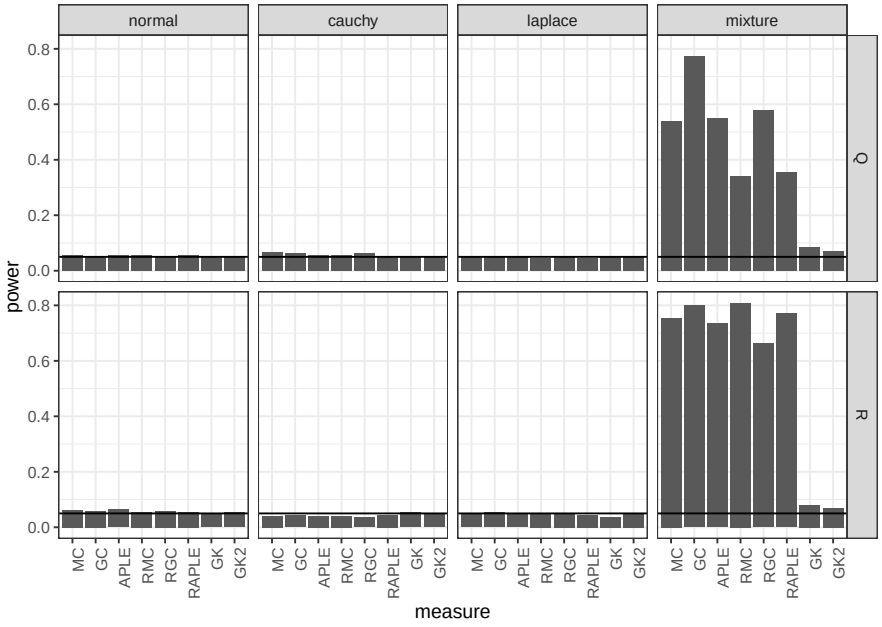

Figure 3: Power comparison for $\rho = 0$, $n = 10$ for Queen (Q) and Rook (R) Weight Matrix.

while GK2 exhibits extreme robustness, our proposed GK measure emerges as the more preferred and balanced choice in terms of robustness.

5 Concluding remarks

In this research, we ventured into the domain of spatial correlation measures, aiming to introduce methods that would robustify traditional metrics. Our approach was grounded in the traditional examination of influence functions, as delineated by (Hampel, 1974). This analysis granted us the opportunity to evaluate the robustness of conventional metrics with our novel propositions.

Leveraging the capabilities of Monte Carlo experiments, we simulated diverse spatial data sets and incorporated varying outlier scenarios. Such simulations were pivotal in discerning the efficacy of our proposed measures. It's worth noting that the realm of robust estimation is deeply intertwined with outlier detection. Our research trajectory subsequently led us to harness our robust measures in devising methods adept at identifying spatial outliers.

For the Normal, Cauchy, and Laplace distributions, the consistent robustness of all measures was evident. In stark contrast, the Mixture distribution presented a more convoluted scenario. However, in the face of such variability, our proposed GK and GK2 metrics stood out as exemplary in their resilience

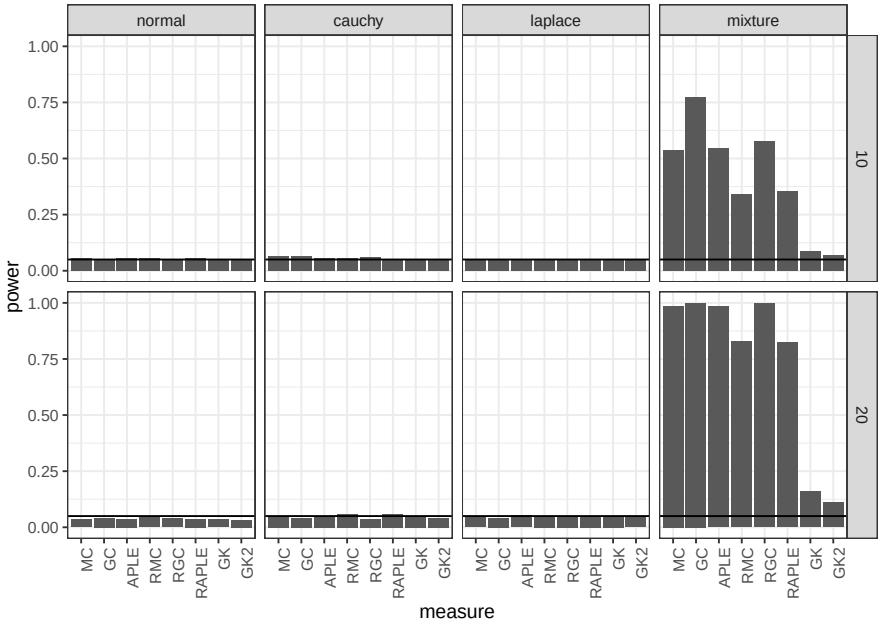

Figure 4: Power comparison for $\rho = 0$, Queen (Q) for $n = 10$ and $n = 20$.

to the extreme values.

In conclusion, the implications of our findings are profound. The GK and GK2 measures, with their inherent robustness, usher in a fresh perspective to spatial data analysis. These findings emphasize the continual need for innovative methods that are both robust and adaptive to the ever-evolving challenges of spatial data.

As we look to the future, our research endeavors will pivot towards the realm of outlier detection. We plan to adapt these measures to their local versions and integrate them into modeling frameworks. Furthermore, this foundational study paves the way for evaluating the influence function across spatial dimensions. Such an approach promises insights into the impact of observed values, taking into consideration not just the magnitude but also the spatial location of the observation. This spatially-aware influence function evaluation has the potential to significantly enrich our understanding of spatial data dynamics.

References

Guiseppe Arbia et al. *Statistical effects of spatial data transformations: a proposed general framework*. Taylor and Francis New York, 1989.

Jørgen Bang-Jensen and Gregory Gutin. Theory, algorithms and applications. *Springer Monographs in Mathematics, Springer-Verlag London Ltd., London*, 101, 2007.

Catherine A Calder and Noel Cressie. Some topics in convolution-based spatial modeling. *Proceedings of the 56th Session of the International Statistics Institute*, pages 22–29, 2007.

Noel Cressie. Aggregation in geostatistical problems. In *Geostatistics Tróia'92: Volume 1*, pages 25–36. Springer, 1993.

Susan J Devlin, Ramanathan Gnanadesikan, and Jon R Kettenring. Robust estimation and outlier detection with correlation coefficients. *Biometrika*, 62 (3):531–545, 1975.

Robert C Geary. The contiguity ratio and statistical mapping. *The incorporated statistician*, 5(3):115–146, 1954.

Daniel A Griffith. The moran coefficient for non-normal data. *Journal of Statistical Planning and Inference*, 140(11):2980–2990, 2010.

Frank R Hampel. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393, 1974.

Hongfei Li, Catherine A Calder, and Noel Cressie. Beyond moran's i: testing for spatial dependence based on the spatial autoregressive model. *Geographical analysis*, 39(4):357–375, 2007.

Hongfei Li, Catherine A Calder, and Noel Cressie. One-step estimation of spatial dependence parameters: Properties and extensions of the aple statistic. *Journal of Multivariate Analysis*, 105(1):68–84, 2012.

Patrick AP Moran. The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):243–251, 1948.

Patrick AP Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37 (1/2):17–23, 1950.

M Hashem Pesaran and Cynthia Fan Yang. Econometric analysis of production networks with dominant units. *Journal of Econometrics*, 219(2):507–541, 2020.

Appendix

Table A.1: Power simulation results for Normal distribution (N=100) and Queen (Q) or Rook (R) W matrices

Measure	n	W	ρ						
			-0.7	-0.5	-0.3	0	0.3	0.5	0.7
MC	10	Q	0.96	0.78	0.43	0.06	0.54	0.90	1.00
MC	10	R	1.00	0.99	0.73	0.06	0.73	0.99	1.00
GC	10	Q	0.55	0.39	0.22	0.05	0.43	0.83	0.99
GC	10	R	1.00	0.99	0.73	0.06	0.73	0.98	1.00
APLE	10	Q	0.96	0.78	0.42	0.06	0.54	0.90	1.00
APLE	10	R	1.00	0.99	0.73	0.06	0.74	0.99	1.00
RMC	10	Q	0.88	0.66	0.34	0.06	0.49	0.88	0.99
RMC	10	R	1.00	0.98	0.68	0.06	0.69	0.98	1.00
RGC	10	Q	0.85	0.62	0.33	0.05	0.48	0.87	0.99
RGC	10	R	1.00	0.96	0.65	0.06	0.66	0.97	1.00
RAPLE	10	Q	0.89	0.68	0.35	0.06	0.49	0.88	0.99
RAPLE	10	R	1.00	0.99	0.69	0.06	0.69	0.98	1.00
GK	10	Q	0.78	0.51	0.25	0.05	0.31	0.68	0.94
GK	10	R	0.99	0.86	0.50	0.05	0.48	0.89	0.99
GK2	10	Q	0.60	0.38	0.21	0.05	0.28	0.58	0.90
GK2	10	R	0.99	0.78	0.44	0.05	0.41	0.81	0.98

Table A.2: Power simulation results for Cauchy distribution (N=100) and Queen (Q) or Rook (R) W matrices

Measure	n	W	ρ						
			-0.7	-0.5	-0.3	0	0.3	0.5	0.7
MC	10	Q	0.99	0.91	0.64	0.06	0.78	0.97	1.00
MC	10	R	1.00	0.99	0.89	0.04	0.89	0.99	1.00
GC	10	Q	0.76	0.61	0.32	0.06	0.47	0.93	0.99
GC	10	R	1.00	0.99	0.84	0.04	0.83	0.99	1.00
APLE	10	Q	0.99	0.91	0.62	0.06	0.78	0.98	1.00
APLE	10	R	1.00	0.99	0.88	0.04	0.89	1.00	1.00
RMC	10	Q	1.00	1.00	0.96	0.06	0.96	1.00	1.00
RMC	10	R	1.00	1.00	0.97	0.04	0.97	1.00	1.00
RGC	10	Q	0.19	0.16	0.11	0.06	0.75	1.00	1.00
RGC	10	R	0.98	0.91	0.70	0.04	0.93	1.00	1.00
RAPLE	10	Q	1.00	0.99	0.96	0.05	0.97	1.00	1.00
RAPLE	10	R	1.00	1.00	0.98	0.04	0.98	1.00	1.00
GK	10	Q	0.99	0.92	0.63	0.05	0.73	0.97	1.00
GK	10	R	1.00	0.99	0.81	0.05	0.79	0.98	1.00
GK2	10	Q	0.63	0.42	0.26	0.05	0.74	0.96	1.00
GK2	10	R	1.00	0.98	0.74	0.05	0.69	0.97	1.00

Table A.3: Power simulation results for Laplace distribution (N=100) and Queen (Q) or Rook (R) W matrices

Measure	n	W	ρ						
			-0.7	-0.5	-0.3	0	0.3	0.5	0.7
MC	10	Q	0.97	0.78	0.43	0.05	0.57	0.90	1.00
MC	10	R	1.00	0.99	0.73	0.05	0.72	0.98	1.00
GC	10	Q	0.57	0.38	0.20	0.05	0.42	0.83	0.99
GC	10	R	1.00	0.99	0.73	0.05	0.72	0.98	1.00
APLE	10	Q	0.96	0.78	0.44	0.05	0.57	0.91	1.00
APLE	10	R	1.00	0.99	0.74	0.05	0.72	0.98	1.00
RMC	10	Q	0.94	0.74	0.42	0.05	0.60	0.92	1.00
RMC	10	R	1.00	0.99	0.73	0.05	0.75	0.99	1.00
RGC	10	Q	0.85	0.63	0.34	0.05	0.54	0.91	1.00
RGC	10	R	1.00	0.97	0.69	0.05	0.70	0.98	1.00
RAPLE	10	Q	0.96	0.78	0.45	0.04	0.61	0.93	1.00
RAPLE	10	R	1.00	0.99	0.75	0.04	0.75	0.99	1.00
GK	10	Q	0.87	0.62	0.31	0.05	0.42	0.81	0.98
GK	10	R	1.00	0.92	0.56	0.04	0.56	0.94	1.00
GK2	10	Q	0.66	0.43	0.24	0.05	0.37	0.73	0.96
GK2	10	R	1.00	0.87	0.50	0.05	0.51	0.87	0.99

Table A.4: Power simulation results for Mixture distribution (N=100) and Queen (Q) or Rook (R) W matrices

Measure	n	W	ρ						
			-0.7	-0.5	-0.3	0	0.3	0.5	0.7
MC	10	Q	0.61	0.19	0.04	0.54	0.95	1.00	1.00
MC	10	R	1.00	0.70	0.10	0.75	1.00	1.00	1.00
GC	10	Q	0.02	0.01	0.00	0.77	0.98	1.00	1.00
GC	10	R	0.99	0.66	0.09	0.80	1.00	1.00	1.00
APLE	10	Q	0.60	0.19	0.04	0.55	0.96	1.00	1.00
APLE	10	R	0.99	0.67	0.08	0.74	1.00	1.00	1.00
RMC	10	Q	0.68	0.38	0.14	0.34	0.86	0.99	1.00
RMC	10	R	0.99	0.64	0.10	0.81	1.00	1.00	1.00
RGC	10	Q	0.29	0.07	0.02	0.58	0.96	1.00	1.00
RGC	10	R	0.98	0.63	0.09	0.66	1.00	1.00	1.00
RAPLE	10	Q	0.68	0.39	0.15	0.36	0.87	0.99	1.00
RAPLE	10	R	0.98	0.59	0.07	0.77	1.00	1.00	1.00
GK	10	Q	0.72	0.41	0.20	0.09	0.47	0.80	0.96
GK	10	R	0.99	0.83	0.38	0.08	0.59	0.93	0.99
GK2	10	Q	0.56	0.32	0.16	0.07	0.35	0.69	0.93
GK2	10	R	0.97	0.74	0.31	0.07	0.49	0.85	0.99

Table A.5: Power simulation results for Normal distribution (N=400) and Queen (Q) W matrix

Measure	n	W	ρ						
			-0.7	-0.5	-0.3	0	0.3	0.5	0.7
MC	20	Q	1.00	1.00	0.91	0.03	0.95	1.00	1.00
GC	20	Q	0.99	0.91	0.56	0.04	0.85	1.00	1.00
APLE	20	Q	1.00	1.00	0.91	0.04	0.96	1.00	1.00
RMC	20	Q	1.00	0.99	0.82	0.04	0.93	1.00	1.00
RGC	20	Q	1.00	0.99	0.80	0.04	0.93	1.00	1.00
RAPLE	20	Q	1.00	0.99	0.82	0.04	0.93	1.00	1.00
GK	20	Q	1.00	0.97	0.67	0.04	0.77	1.00	1.00
GK2	20	Q	0.98	0.86	0.49	0.03	0.69	0.98	1.00

Table A.6: Power simulation results for Cauchy distribution (N=400) and Queen (Q) W matrix

Measure	n	W	ρ						
			-0.7	-0.5	-0.3	0	0.3	0.5	0.7
MC	20	Q	1.00	1.00	0.98	0.05	0.98	1.00	1.00
GC	20	Q	0.99	0.97	0.83	0.04	0.96	1.00	1.00
APLE	20	Q	1.00	1.00	0.98	0.05	0.98	1.00	1.00
RMC	20	Q	1.00	1.00	1.00	0.06	1.00	1.00	1.00
RGC	20	Q	0.13	0.11	0.09	0.04	1.00	1.00	1.00
RAPLE	20	Q	1.00	1.00	1.00	0.06	1.00	1.00	1.00
GK	20	Q	1.00	1.00	0.98	0.05	1.00	1.00	1.00
GK2	20	Q	0.98	0.85	0.52	0.04	1.00	1.00	1.00

Table A.7: Power simulation results for Laplace distribution (N=400) and Queen (Q) W matrix

Measure	n	W	ρ						
			-0.7	-0.5	-0.3	0	0.3	0.5	0.7
MC	20	Q	1.00	1.00	0.92	0.05	0.95	1.00	1.00
GC	20	Q	0.99	0.91	0.60	0.04	0.84	1.00	1.00
APLE	20	Q	1.00	1.00	0.92	0.05	0.95	1.00	1.00
RMC	20	Q	1.00	1.00	0.92	0.05	0.97	1.00	1.00
RGC	20	Q	1.00	0.99	0.84	0.05	0.95	1.00	1.00
RAPLE	20	Q	1.00	1.00	0.92	0.05	0.97	1.00	1.00
GK	20	Q	1.00	0.99	0.78	0.05	0.87	1.00	1.00
GK2	20	Q	0.99	0.92	0.62	0.05	0.81	1.00	1.00

Table A.8: Power simulation results for Mixture distribution (N=400) and Queen (Q) W matrix

Measure	n	W	ρ						
			-0.7	-0.5	-0.3	0	0.3	0.5	0.7
MC	20	Q	0.96	0.39	0.01	0.99	1.00	1.00	1.00
GC	20	Q	0.00	0.00	0.00	1.00	1.00	1.00	1.00
APLE	20	Q	0.96	0.37	0.01	0.98	1.00	1.00	1.00
RMC	20	Q	0.98	0.73	0.20	0.83	1.00	1.00	1.00
RGC	20	Q	0.64	0.11	0.00	1.00	1.00	1.00	1.00
RAPLE	20	Q	0.98	0.73	0.20	0.82	1.00	1.00	1.00
GK	20	Q	1.00	0.88	0.47	0.16	0.90	1.00	1.00
GK2	20	Q	0.98	0.80	0.44	0.11	0.81	1.00	1.00