Mathematical Geosciences (2024) 56:103–119
https://doi.org/10.1007/s11004-023-10080-y

SPECIAL ISSUE

EpiGeostats: An R Package to Facilitate Visualization of Geostatistical Disease Risk Maps

Manuel Ribeiro[1] · **Leonardo Azevedo**[1] · **Maria João Pereira**[1]

Received: 16 January 2023 / Accepted: 11 June 2023 / Published online: 24 July 2023

Abstract

With the emergence of the coronavirus disease 2019 (COVID-19) pandemic in Portugal, a geostatistical tool was developed to model the spatial distribution of COVID-19 risk to support decision-making and policymakers. Based on a block direct sequential simulation algorithm, the model provides detailed disease risk estimates and associated spatial uncertainty. However, uncertainty is difficult to visualize with the estimated risk, and is usually overlooked as a tool to support decision-making. Ignoring uncertainty can be misleading in evaluating risk, since the amount of uncertainty varies throughout the spatial domain. The EpiGeostats R package was developed to solve this problem, since it integrates the geostatistical model and visualization tools to deliver a single map summarizing disease risk and spatial uncertainty. This paper briefly describes the methodology and package functions implemented for interfacing with the tools in question. The use of EpiGeostats is illustrated by applying it to real data from COVID-19 incidence rates on mainland Portugal. EpiGeostats is a powerful tool for supporting decision-making in the context of epidemics, since it combines a well-established geostatistical model for disease risk mapping with simple and intuitive ways of visualizing results, which prevent fine-scale inference in regions with high-risk uncertainty. The package may be used for similar problems such as mortality risk, or applied to other fields such as ecology or environmental epidemiology.

Keywords Disease mapping · R package · Block direct sequential simulation · Map visualization · Spatial uncertainty

✉ Manuel Ribeiro
manuel.ribeiro@tecnico.ulisboa.pt

[1] CERENA, DER, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisbon, Portugal

1 Introduction

Making decisions to mitigate the impact of the coronavirus disease 2019 (COVID-19) pandemic is a highly difficult and complex task requiring policymakers to compile and analyse relevant information in a timely and effective manner. Among the array of information required, geographical distribution of disease plays a central role, since it helps to understand the spatial clustering and transmission trends of ongoing COVID-19 outbreaks (Ahasan and Hossain 2021).

A type of map frequently used for visualizing spatial distribution of COVID-19 risk are choropleth maps, which use different colour and pattern combinations to represent counts (or rates) of aggregated data by region or area (e.g. administrative boundaries). At the national or local level, these types of maps helped health authorities to model and predict the spread of COVID-19 disease (Giuliani et al. 2020), visualize the effects of pandemic lockdowns (Carroll and Prentice 2021) and detect COVID-19 'hotspots' (Guillette et al. 2020). At the same time, choropleth maps show constant risk per region with sudden discontinuities at the region's borders, disregarding that disease risk varies continuously across spatial domains. For this reason, choropleth maps are considered a relatively crude method of displaying disease data (Waller and Gotway 2004). Isopleth risk maps are an alternative way to visualize the spatial distribution of a disease from aggregated data, since they provide a spatially continuous disease risk over the spatial domain (Goovaerts 2005; Jaya and Folmer 2020).

Azevedo et al. (2020) proposed an isopleth COVID-19 risk map to inform and support decision-makers in monitoring and evaluating spread dynamics on mainland Portugal. The approach, based on geostatistical methods and stochastic simulation algorithms, produces highly resolved disease maps and addresses the biased visual perception produced with choropleth maps by imposing a spatially continuous variability with no sharp discontinuities at the boundaries of administrative regions. In addition, the proposed approach accounts for the noise attached to risk estimated from small population sizes and the spatial uncertainty associated with predicting risk at unmonitored locations.

The model relies on the assumption that COVID-19 risk obtained from aggregated data involves observations derived from an underlying latent, spatially varying fine-spatial-resolution risk field, and a spatial covariance model inferred from coarse administrative-level observed disease incidence data weighted by population size. It is then incorporated into a flexible geostatistical simulation algorithm, block direct sequential simulation (Rita et al. 2013) that considers the varying size and shape of administrative regions to generate COVID-19 risk on a gridded surface from data available only at areal level. In fact, the flexibility of this algorithm relies exactly on the ability to deal with both point data (i.e. predicting disease risk at each grid node) and areal data (i.e. disease rates observed in administrative regions) accommodating the change in the size and shape of administrative regions across the spatial domain. As any stochastic sequential simulation algorithm, each run (i.e. stochastic realization) of the block direct sequential simulation produces a different map (see Fig. 1), since the conditioning data at each location varies due to the random path used to simulate each pixel sequentially (Deutsch and Journel 1992). Azevedo et al. (2020) present in

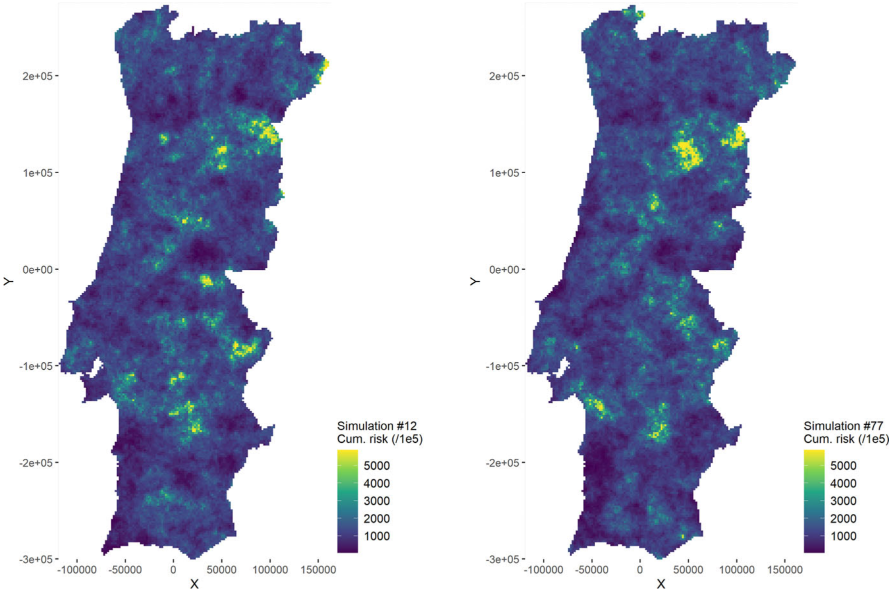


Fig. 1 Two stochastic realizations obtained with block direct sequential simulation illustrate the variability generated by the algorithm. Any of these maps captures the major patterns of spatial covariance and other statistical parameters (e.g. the mean or variance) estimated from observed data

detail the description of the block direct sequential algorithm and its application to model the spatial incidence risk for COVID-19 on mainland Portugal.

As a result, two critical maps for pandemic response can be drawn from the set of simulations (Fig. 2): one presenting the median COVID-19 incidence risk, derived from the point-wise median computed from the set of simulations, from which it is possible to identify areas where high disease risk is expected; and the other with the associated spatial uncertainty as revealed by the point-wise variance or interquartile range computed from the same set of simulations.

Presenting the pair of maps side by side (as shown in Fig. 2) provides policymakers with critical data to identify areas of high/low risk with high spatial uncertainty and areas of high/low risk with low spatial uncertainty. However, in practice, the two are rarely combined, because uncertainty maps are usually relegated as a matter of secondary importance or are simply ignored due to the difficulty of visual comparison across multiple maps (Taylor et al. 2020).

To overcome this problem, Taylor et al. (2020) proposed pixelation to simultaneously visualize disease risk and uncertainty in a single map, and illustrated the methodology to map 2017 *Plasmodium falciparum* incidence in central Africa as proof of concept. The proposed solution provides disease risk maps with varying pixel size so that areas of high average uncertainty have large pixels, while areas with low average uncertainty have small pixels. This visualization approach is a valuable tool to rapidly identify high risk in areas with high spatial uncertainty and high risk in areas with low spatial uncertainty. Currently, an R package performing pixelation

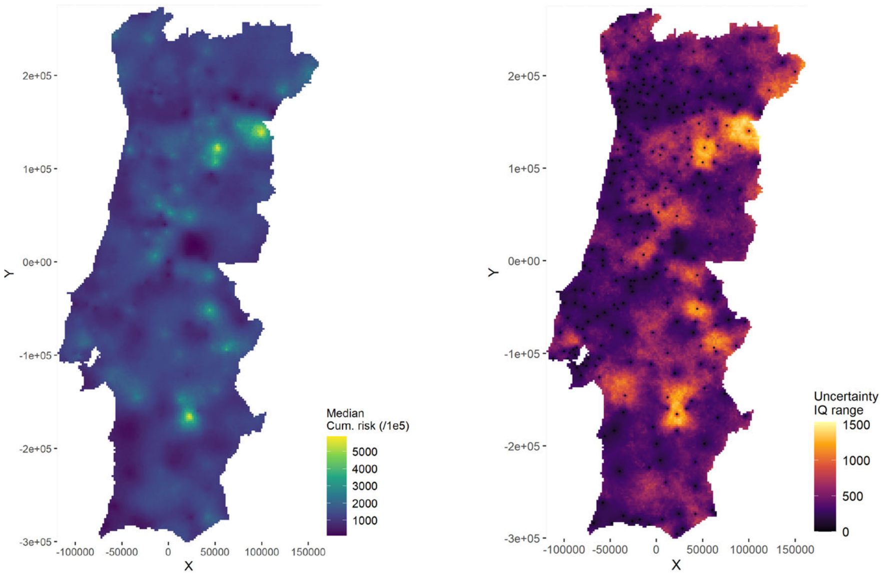


Fig. 2 Maps illustrating the median COVID-19 risk on mainland Portugal on 15 January 2021 (left); and the interquartile distance (right) computed from a set of 100 simulations using the algorithm proposed by Azevedo et al. (2020)

computations for map visualization is available online (https://github.com/artaylor85/pixelate).

While the block direct sequential simulation algorithm is implemented as a standalone software tool distributed free of charge, it is not open source, and it can be cumbersome to use. Therefore, software solutions combining disease risk mapping based on the modelling approach proposed by Azevedo et al. (2020) with the pixelation approach to visualize disease risk and associated uncertainty in a single map are not available, and require a considerable programming effort that is prone to human error. To bridge this gap, we developed the EpiGeostats R package for the rapid computation and visualization of disease risk and uncertainty in a single map, based on the geostatistical disease modelling approach proposed by Azevedo et al. (2020) and the visualization solution proposed by Taylor et al. (2020).

The target audiences of the EpiGeostats package are policymakers and researchers interested in disease risk mapping, modelling and communication. The EpiGeostats package is flexible, given that users may understand, modify or develop the code for their own work, and can be adapted to other transmissible diseases and similar problems such as mortality risk, or applied to other fields such as ecology, criminology or environmental epidemiology.

2 Implementation

The EpiGeostats package was developed with R software (R Core Team 2021), which is a free, open-source environment for statistical and scientific computing. The package runs in Windows and is freely available to users from the GitHub platform (https://github.com/maluicr/EpiGeostats). To use EpiGeostats, dss.c.64.exe software (https://github.com/maluicr/dss) must be downloaded to perform the block direct sequential simulation algorithm and to install the pixelate R package (https://github.com/aimeertaylor/pixelate) to enhance visualization of disease risk map results. To properly execute all of these steps, the annotated R code is available online at the GitHub repository (https://github.com/maluicr/EpiGeostats) in the readme.md file (in the root directory of the repository).

Throughout this section, the tools, methods, input data and functions available in the package are described.

2.1 Stand-Alone Software dss.c.64.exe

The executable dss.c.64.exe is a free and open-source set of geostatistical simulation methods for modelling natural and environmental phenomena. It was developed by researchers at CERENA [Centro de Recursos Naturais e Ambiente] (https://cerena.pt/) and is one of the few types of software that provides the block direct sequential simulation algorithm. The EpiGeostats package provides convenient wrappers to read datasets and write files in formats readable by dss.c.64.exe (Geo-EAS file format, Deutsch and Journel 1992) and invokes dss.c.64.exe for running the block direct sequential simulation algorithm.

To use dss.c.64.exe with EpiGeostats, the following steps are required: (1) create a folder named ‘input’ in the working directory, and (2) download dss.c.64.exe to that folder.

2.2 The Pixelate R Package

The package pixelate proposes an elegant solution to the problem of visualizing spatial uncertainty in geostatistical maps. Specifically, it provides a single function to compute a map of varying pixel size, where areas with a higher density of pixels (i.e. with small pixel size) represent lower average uncertainty areas, and areas with a lower density of pixels (i.e. large pixel size) indicate higher average uncertainty areas, using a rationale analogous to highly versus lowly resolved satellite images (Taylor et al. 2020). Moreover, the function calls other functions from the ggplot2 package (Wickham 2016) to allow the visualization of elegant output maps.

2.3 Geostatistical Framework

Block sequential simulation, implemented in dss.c.64.exe software, was developed to make predictions and quantify their spatial uncertainty in cases where combining spatial data with different spatial units (also known as spatial support) is required. Among geostatisticians, this is known as a ‘change-of-support problem’, and has been described extensively in the literature (e.g. Emery 2009; Goovaerts 2005; Journel and Huijbregts 1978; Kyriakidis 2004; Meng et al. 2019; Young et al. 2009; Young and Gotway 2010; Zaytsev et al. 2016).

The EpiGeostats package focuses on the specific case of disease mapping as presented by Azevedo et al. (2020), where the goal is to provide a tool predicting disease risk with high resolution in a spatially continuous domain (here, the spatial support is point data) from data only available at an areal level with varying spatial sizes and shapes (i.e. the spatial support is block data). Within this framework, point data are represented by a rectangular grid of nodes with regular spacing partitioned into a finite number of regions (i.e. municipalities), and block data refer to measurements observed in those regions (i.e. aggregated number of disease cases per unit of time, assigned to each municipality). The block direct sequential simulation algorithm provides the means to generate stochastic realizations (or simulated maps) with high resolution, reproducing disease incidence fluctuations observed in block data with similar statistical properties (i.e. empirical histogram and spatial covariance) (Gómez-Hernández and Srivastava 2021). The uncertainty of incidence rates derived from small population sizes (a problem commonly known as the ‘small number problem’) is accounted for using Poisson kriging (Goovaerts 2005, 2008).

In a nutshell, the block direct sequential simulation algorithm proceeds as follows:

1. It randomly defines a path over the entire spatial domain, passing through all u_i grid nodes to be simulated ($i = 1, \ldots, N$).
2. For node u_i, it searches the conditioning data (neighbour point data, previously simulated values and block data).
3. It uses the conditioning data to derive local covariance values: block-to-block, block-to-point, point-to-block and point-to-point.
4. It builds and solves the block kriging system and obtains the local mean and variance estimates at node u_i.
5. It draws a simulated value from the global probability distribution function centred on the local mean and delimited by the local variance obtained in step 4.
6. It adds the simulated value to the dataset and repeats steps 2 to 5 for u_{i+1} until all grid nodes have been simulated for one geostatistical realization.

A set of realizations is obtained after repeating steps 1 to 6 until a given predefined number of realizations are generated, representing a set of ‘images’ with a continuous spatial surface reproducing (on average) the statistical properties of observed block data as revealed by the spatial covariance and the empirical histogram. Two key maps may be drawn from the set of simulations: a point-wise median map of the variable of interest and the attached spatial uncertainty, which is quantified by the point-wise interquartile range (IQR).

In the EpiGeostats package, the covariance of the disease risk, or equivalently its semi-variogram, is required to derive the kriging weights and kriging variance referred in step 4 of the algorithm. It is worth noting that the function implemented in EpiGeostats to fit the semi-variogram is population-weighted, as presented in Azevedo et al. (2020), consequently mitigating the impacts of having regions with varying population sizes in disease risk uncertainty (Goovaerts 2006; Young and Gotway 2010).

2.4 Tool Development

The EpiGeostats package requires two input data files, where one contains geographic information about the partition of the spatial domain into regions (e.g. municipalities or counties), and the other contains data on the aggregated number of disease cases observed in each region. Then, using EpiGeostats function calls, users must set parameters required for geostatistical simulations and write the input files in readable format for dss.c.64.exe, after which the block direct sequential simulation algorithm can be executed. The results include a set of simulated disease risk maps, a median disease risk map, a spatial risk uncertainty map and a pixelated map representing both disease risk and spatial risk uncertainty in a single map, which can then be extracted and/or plotted in R. Figure 3 represents a schematic overview of the EpiGeostats R package.

The package is accompanied by detailed documentation (see Sect. 2.5) and examples for easy replication.

2.4.1 Input Data

The input required for the EpiGeostats R package is comprised of two data frames: (i) block data, a data frame with an aggregated number of incidence cases per area or subregion, and (ii) point data, a data frame representing a regular grid dataset with id subregion values at each grid node. The data frame structure should follow these specifications:

(i) Block data are a data frame with geographic coordinates x, y, region id, number of disease cases and size of population (i.e. population at risk). A z-coordinate column filled in with a constant integer (e.g. 1) should also be included, only to match the file structure required for dss.c.64.exe. No blank values are allowed.

(ii) Point data refer to a rectangular grid of nodes with regular spacing covering the entire aerial extent to be modelled. The data frame needs to include region id values at all simulation x, y coordinates (i.e. grid nodes). All regions represented in disease (block) data should be represented by one or more grid nodes in point data.

2.4.2 Functions

The EpiGeostats package includes nine functions that are briefly summarized in Table 1. Additional details about the functions can be found in the package vignette or by accessing their documentation pages with the help() function (or ? operator) inside the R session.

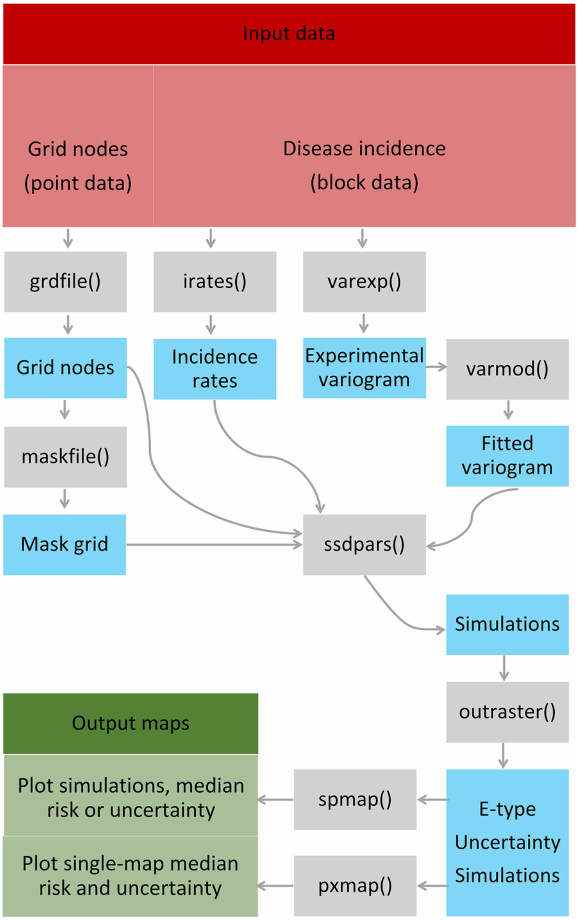


Fig. 3 Schematic overview of the features of the EpiGeostats package. The red box represents the input data; the green box represents output results that can be plotted; the blue boxes represent results extracted with EpiGeostats functions, represented by grey boxes; and the arrow lines represent relations between features

The collection of files read and created by dss.c.64.exe follows the filename convention [YYYYMMDD]_[dataset].[*], where [YYYYMMDD] is the date specified in argument 'day' of irates(), [dataset] refers to the object printed to the file and [*] refers to the filename extension.

As an example, consider the COVID-19 incidence on mainland Portugal on 15 January 2021 (see example in Sect. 3). Table 2 presents the names of the files generated in this example with EpiGeostats functions.

All files are written and stored in the 'input' folder.

Table 1 Brief description of R functions implemented in the EpiGeostats package

Function	Description
irates()	Reads block data frame, computes disease rates and writes a file with results in a format readable by dss.c.64.exe software
grdfile()	Reads gridded data—a SpatialPixelDataFrame()—and writes a file in a format readable by dss.c.64.exe
maskfile()	Reads the output of grdfile() and writes a file with mask data in a format readable by dss.c.64.exe software
ssdpars()	Writes the parameters file and invokes dss.c.64.exe to run block direct sequential simulation. Generates simulated maps in native file format
varexp()	Computes population-weighted experimental semi-variograms from irates() output. For now, only the omnidirectional case is implemented
varmodel()	Fits (manually) a theoretical semi-variogram. For now, only spherical and exponential models are available
outraster()	Reads simulation files returned by ssdpars(), and returns a list with simulated maps, median e-type and uncertainty maps. Writes grid files (.gri/.grd) with the results
spmap()	Wraps functions from the ggplot2 package to plot more elegant simulation maps, a median risk map or a spatial risk uncertainty map
pxmap()	Wrapper function calling pixelate package to plot a disease risk map with a visual representation of spatial risk uncertainty, as a function of pixel size

Table 2 Example of filenames created with the EpiGeostats package

Function call	R object created	Filename
irates()	Incidence rates	20210115_rate.out
maskfile()	Grid mask	20210115_mask.out
grdfile()	Regular grid	20210115_grid.out
ssdpars()	Parameters file	20210115_ssdr.par
ssdpars()	Set of simulations	20210115_sim_89.out (e.g. simulation #89)
outraster()	Median map	20210115_medn.gri (or. grd)
outraster()	Uncertainty map	20210115_uncr.gri (or. grd)

2.5 Available Documentation

After the EpiGeostats package has been installed, a tutorial to demonstrate its functionalities can be found in the package vignette. The vignette, which can be found by typing help(package = 'EpiGeostats') in the console, provides information through examples, with fully annotated code, on how to put together several functions in order to obtain a specific result.

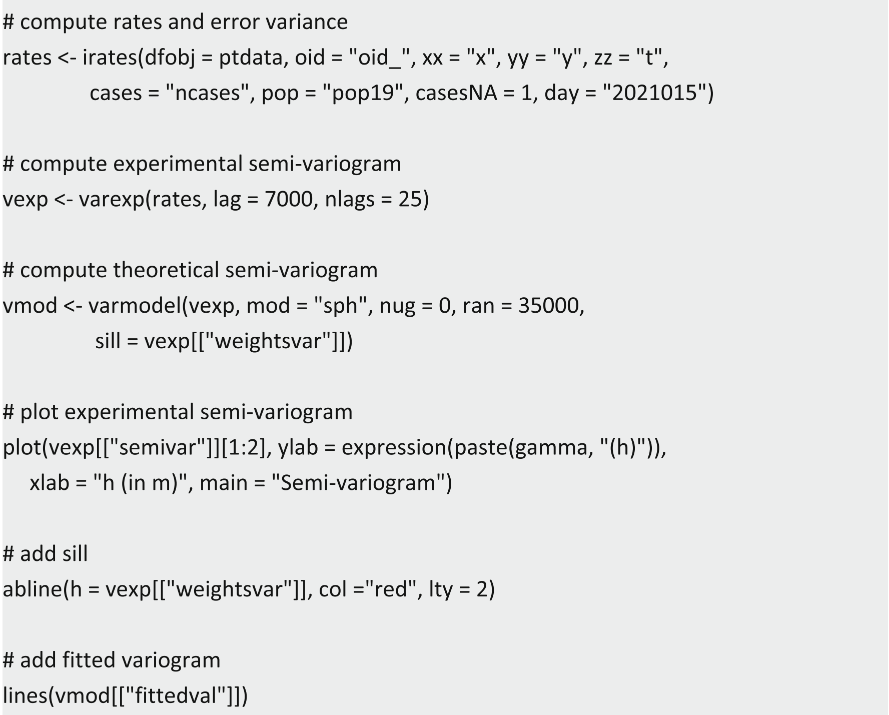
The vignette can also be used as a tutorial to show the role of each function and may serve as a starting point for users to customize different setup choices. Technical documentation with details about the functions is available in the documentation pages accessible using the help() function or ? operator (e.g. ?irates) in the R console. The details include a short description of the underlying function, the calling syntax, function arguments, returned values and other outputs.

3 Results

As an example of application, we next show how the EpiGeostats package can perform disease mapping of COVID-19 incidence on mainland Portugal on 15 January 2021. The COVID-19 dataset ptdata and the regular grid ptgrid comes with the EpiGeostats package. The coordinate reference system is ETRS89/Portugal TM06 (EPSG: 3763) and coordinates are in metres.

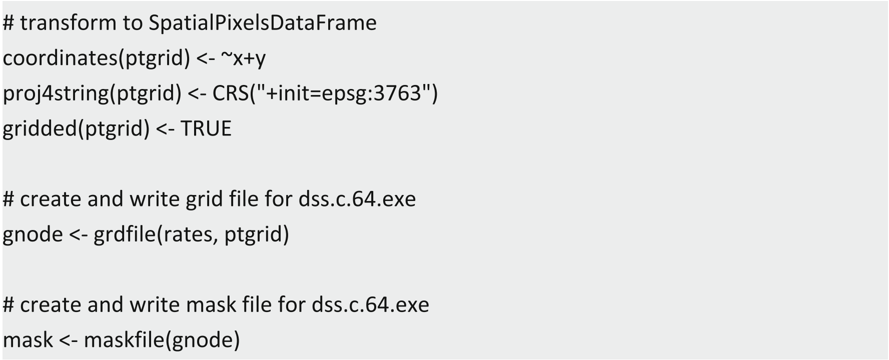
3.1 Running irates(), varexp() and varmodel()

Computing incidence rates and fitting the semi-variogram model requires several manual steps that can be achieved with little R code effort and simple syntax. In Fig. 4, the R code uses the EpiGeostats functions irates() to compute COVID-19 incidence rates



```
# compute rates and error variance
rates <- irates(dfobj = ptdata, oid = "oid_", xx = "x", yy = "y", zz = "t",
                cases = "ncases", pop = "pop19", casesNA = 1, day = "2021015")

# compute experimental semi-variogram
vexp <- varexp(rates, lag = 7000, nlags = 25)

# compute theoretical semi-variogram
vmod <- varmodel(vexp, mod = "sph", nug = 0, ran = 35000,
                 sill = vexp[["weightsvar"]])

# plot experimental semi-variogram
plot(vexp[["semivar"]][1:2], ylab = expression(paste(gamma, "(h)")),
     xlab = "h (in m)", main = "Semi-variogram")

# add sill
abline(h = vexp[["weightsvar"]], col ="red", lty = 2)

# add fitted variogram
lines(vmod[["fittedval"]])
```

Fig. 4 Example of workflow script for the calculations of incidence rates, estimation of semi-variogram and the fit of theoretical semi-variogram



```
# transform to SpatialPixelsDataFrame
coordinates(ptgrid) <- ~x+y
proj4string(ptgrid) <- CRS("+init=epsg:3763")
gridded(ptgrid) <- TRUE

# create and write grid file for dss.c.64.exe
gnode <- grdfile(rates, ptgrid)

# create and write mask file for dss.c.64.exe
mask <- maskfile(gnode)
```

Fig. 5 Transforming grid data into SpatialPixelsDataFrame and generating point and mask data files using grdfile() and maskfile() functions

and varexp() and varmodel() to estimate the population's weighted semi-variograms and fit the theoretical semi-variogram model, respectively.

The function irates() also writes a file with incidence results in a format readable by dss.c.64.exe software. The arguments for the variogram model—varmod()—are set manually by the user in EpiGeostats. These can be obtained by a visual inspection of the semi-variogram estimates. The plot of semi-variogram results (experimental and/or theoretical) can be visualized using the generic R function plot(), as shown in Fig. 4.

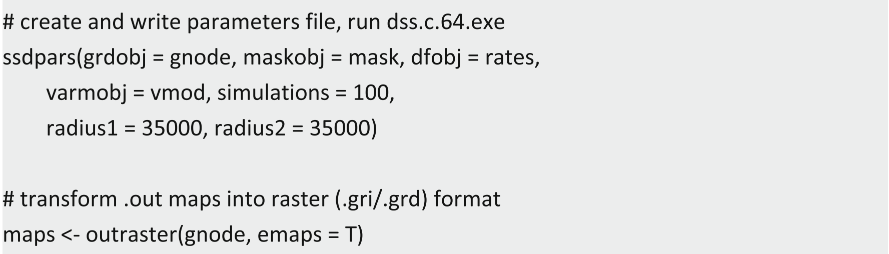
3.2 Running grdfile() and maskfile()

These two functions are designed to convert point data into the native file structure of dss.c.64.exe—grdfile()—and to generate a mask data file—maskfile(). The output files are required by the executable to run block direct sequential simulation.

Running grdfile() requires a data object with spatial attributes that have spatial locations on a grid with regular spacing (point data) and a disease rates object (block data). Users may create a SpatialPixelsDataFrame object using R package gstat (Pebesma and Wesseling 1998) to form a grid with the specified requirements (Fig. 5). The *maskfile()* function is applied to the object output of the function *grdfile()*.

3.3 Running ssdpars() and outraster()

Users must call ssdpars() to run a block direct sequential simulation. This function provides functionalities to write a parameters file to be read by dss.c.64.exe and a shell command invoking the executable to compute block simulations. The parameters file includes the required information to set search grid parameters, variogram model parameters and kriging estimator specifications (Fig. 6). Upon execution, the output data (stochastic simulations) are written to text files in native format (.out) and stored inside the 'input' data directory.



```
# create and write parameters file, run dss.c.64.exe
ssdpars(grdobj = gnode, maskobj = mask, dfobj = rates,
      varmobj = vmod, simulations = 100,
      radius1 = 35000, radius2 = 35000)

# transform .out maps into raster (.gri/.grd) format
maps <- outraster(gnode, emaps = T)
```

Fig. 6 In the R code example, the function ssdpars() runs the block direct sequential simulation algorithm and generates 100 grid simulations. The function outraster() is specified to return the median e-type disease risk map and the spatial risk uncertainty map in raster format

The function outraster() reads output files (grid simulations in .out file format) and writes files in native raster file format (.grd and .gri). Depending on argument specifications, outraster() may return, in raster format, all simulated maps and/or the median e-type disease risk map and the spatial risk uncertainty map.

3.4 Running spmap() and pxmap()

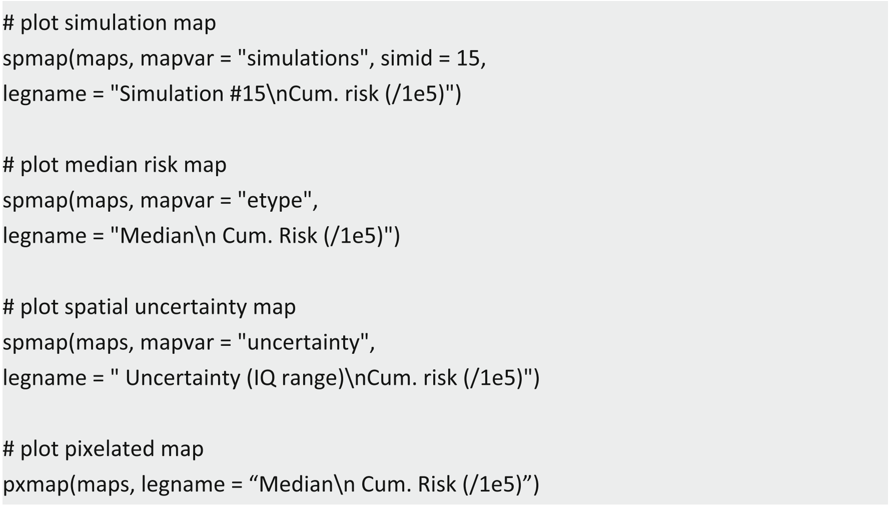
The functions spmap() and pxmap() are used to plot stochastic simulations, the median disease risk map, the spatial risk uncertainty map and the single-map version (a pixelated version) combining disease risk with spatial uncertainty as a function of pixel size (Fig. 7).

Assuming that the median disease risk and the spatial risk uncertainty maps are extracted from simulations with the outraster() function (emaps argument must be set to TRUE), they can be plotted using the spmap() call, a wrapper around ggplot() from the package ggplot2 (Wickham 2016). The pixelated version combining both maps is performed with the R package pixelate (Taylor et al. 2020) and can be plotted with the pxmap() call.

3.5 Output Plots

Figure 8 shows four maps generated with the plotting function calls presented in Fig. 7.

The arguments specified by the pxmap() function to generate the pixelated map can be modified to adjust map visualization to user preferences. Please refer to relevant discussions on map visualization interpretations (Lucchesi and Wikle 2017; Taylor et al. 2020).



```
# plot simulation map
spmap(maps, mapvar = "simulations", simid = 15,
legname = "Simulation #15\nCum. risk (/1e5)")

# plot median risk map
spmap(maps, mapvar = "etype",
legname = "Median\n Cum. Risk (/1e5)")

# plot spatial uncertainty map
spmap(maps, mapvar = "uncertainty",
legname = " Uncertainty (IQ range)\nCum. risk (/1e5)")

# plot pixelated map
pxmap(maps, legname = “Median\n Cum. Risk (/1e5)”)
```

Fig. 7 The functions used to plot the maps require minimum user input of R code with a simple syntax. The function spmap() calls ggplot2 functions to plot simulations, median risk and spatial uncertainty risk maps, while pxmap() adds the function *pixelate()* to the call. Additional arguments are available for map visualization and pixelation

4 Discussion

The EpiGeostats R package resulted from the need to develop a simple tool for rapid computation and visualization of disease risk and spatial uncertainty in a single map, based on the COVID-19 geostatistical modelling approach proposed by Azevedo et al. (2020). To the best of our knowledge, no R implementation of block direct sequential simulation applied to spatial epidemiology yet exists.

The new tool extends the scope of models available in R to map disease risk and support decision-making in the context of epidemics. First of all, the model implemented not only addresses the need of discrete Poisson distribution to map disease counts (or rates) associated to areal regions, but also considers their different geometric sizes, shapes and orientations. This is achieved by discretizing the areal data into a highly resolved and regularly spaced grid of nodes and solving the block kriging equation system for each node. Secondly, the model incorporates an uncertainty analysis derived from an ensemble of maps (stochastic simulations), from which a probability distribution of disease risk values on each grid node is drawn. Finally, to present the results, EpiGeostats borrows the concept of pixelation to simultaneously visualize uncertainty and disease risk in a single map, allowing policy makers to make better use of spatial uncertainty when evaluating risk. Figure 8 (bottom right) illustrates the results of integration, and shows how this package can contribute to more informed decision-making, since it proposes an adequate geostatistical model in the spatial epidemiology context, and rapidly identifies high risk in areas with high spatial uncertainty, and high risk in areas with low spatial uncertainty, thereby preventing fine-scale inference in regions with high-risk uncertainty.

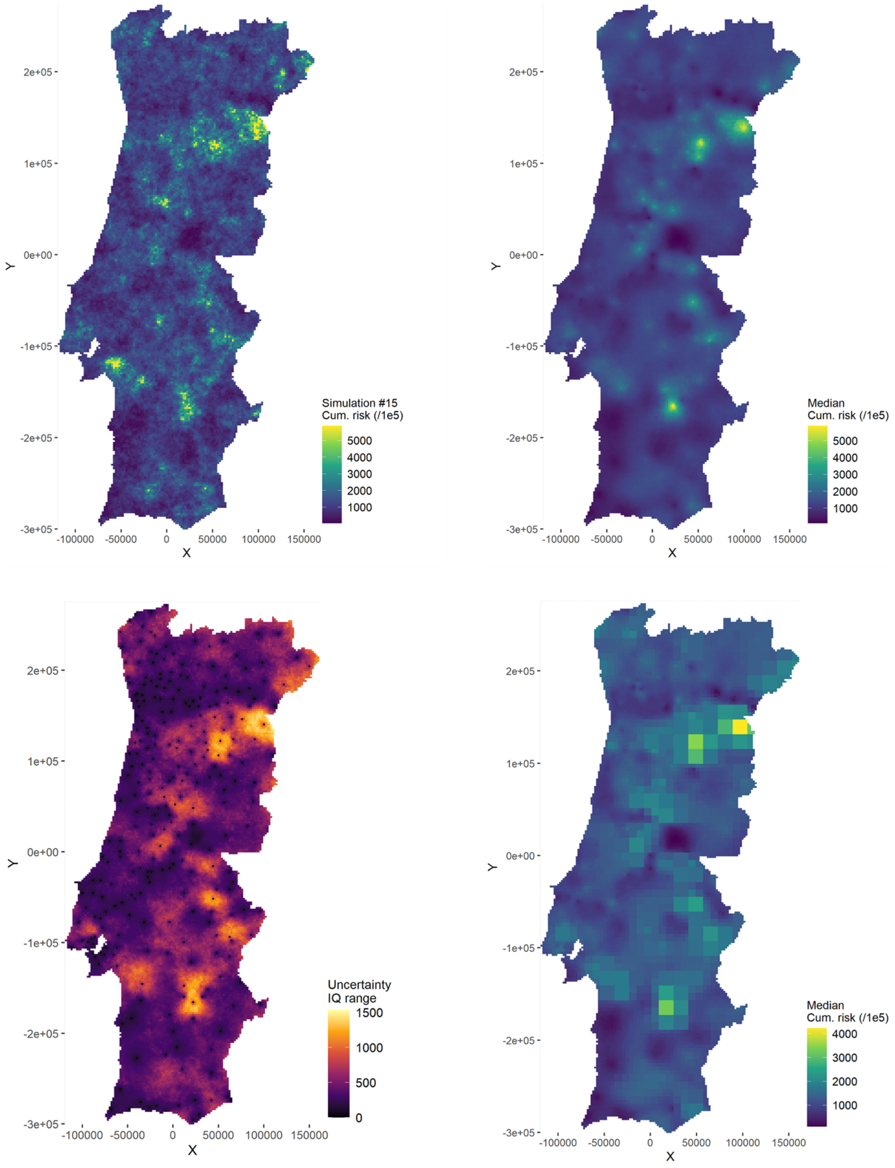


Fig. 8 One block direct sequential simulation (top left), median cumulative disease risk map (top right), spatial risk uncertainty map (bottom left) and pixelated map (bottom right) generated with EpiGeostats functions *spmap()* and *pxmap()*

One limitation of EpiGeostats is that it requires computationally intensive calculations due to inverse matrix operation in solving the block kriging system, which can significantly limit its application if the number of nodes and number of simulations specified is large (e.g. executing 100 simulations on a regular grid with 141 × 288 nodes takes 133 s on an AMD Ryzen 5 5600X 6-core processor 3.70 GHz). While ready on the user end, and parallelized at the CPU thread level, the efficiency of the code will be improved in future versions. Another consideration is related to the fact that experimental population-weighted semi-variograms implemented are the simplest ones, since they are a scalar function of lag (they are omnidirectional). While these capture the main spatial covariance features present in disease data, further developments will provide directional semi-variograms to capture variations in different directions. In addition, only two admissible semi-variogram models have been implemented. They are, however, frequently applied, since they cover the major spatial covariance shapes observed in spatial epidemiology (Goovaerts et al. 2005; Ribeiro and Pereira 2018).

5 Conclusions

EpiGeostats is an R package for disease risk mapping that allows for an uncertainty analysis. The target audiences are policy makers and researchers interested in spatial modelling of disease and data visualization. The package allows for more informed decision-making in the context of epidemics. Package functions require minimum user input of R code and are accompanied by detailed documentation and workflow examples. Future extensions may include the development of a user-friendly, web-based Shiny app or the implementation of more flexible geostatistical models (e.g. adding potential covariates). Moreover, EpiGeostats can be easily extended to other fields, such as ecology or environmental epidemiology.

Acknowledgements The data related to COVID-19 cases presented in this manuscript were provided by the Portuguese Directorate-General of Health (DGS). The authors would like to thank Alzira Ramos, António Rodríguez-Lizana and the two anonymous reviewers for their helpful comments and suggestions.

Authors' Contributions MR contributed to the conception, software development, writing the original draft, review and editing, visualization, supervision and validation. LA contributed to the conception, review and editing and software development. MJP contributed to the conception and review and editing. All authors have read and approved the manuscript.

Funding Open access funding provided by FCT|FCCN (b-on). This work was supported by CERENA—-Centro de Recursos Naturais e Ambiente—research centre (Grant FCT-UIDB/04028/2020). The project SCOPE—Spatial Data Sciences for COVID-19 Pandemic (Grant DSAIPA/DS/0115/2020) provided the open-access funding. MR is supported by the Fundação para a Ciência e Tecnologia with a research contract IF2018/CP1384/IST-ID/175/2018. The funding sources had no role in the design, implementation, analysis or interpretation of the data, or the decision to submit the results.

Availability of Data and Materials The dataset used in the manuscript is available in the EpiGeostats package. EpiGeostats and dss.c.64.exe for Windows (64 bits) are available via GitHub at the repositories https://github.com/maluicr/EpiGeostats and https://github.com/maluicr/dss, respectively. The EpiGeostats repository includes annotated R code demonstrating how to install and run the package and download the executable (see the readme.md file in the root of repository). After the package has been installed, a tutorial with examples to demonstrate its functionalities can be found in the package vignette.

Declarations

Conflict of interest The authors declare that they have no competing interests.

References

Ahasan R, Hossain MM (2021) Leveraging GIS and spatial analysis for informed decision-making in COVID-19 pandemic. Health Policy Technol 10:7–9. https://doi.org/10.1016/j.hlpt.2020.11.009

Azevedo L, Pereira MJ, Ribeiro MC, Soares A (2020) Geostatistical COVID-19 infection risk maps for Portugal. Int J Health Geogr 19:1–8. https://doi.org/10.1186/s12942-020-00221-5

Carroll R, Prentice CR (2021) Using spatial and temporal modelling to visualize the effects of U.S. state issued stay at home orders on COVID-19. Sci Rep 11:1–7. https://doi.org/10.1038/s41598-021-93433-z

Deutsch CV, Journel AG (1992) GSLIB: geostatistical software library and user's guide. Oxford University Press, New York

Emery X (2009) Change-of-support models and computer programs for direct block-support simulation. Comput Geosci 35:2047–2056. https://doi.org/10.1016/j.cageo.2008.12.010

Giuliani D, Dickson MM, Espa G, Santi F (2020) Modelling and predicting the spatio-temporal spread of cOVID-19 in Italy. BMC Infect Dis 20:1–10. https://doi.org/10.1186/s12879-020-05415-7

Gómez-Hernández JJ, Srivastava RM (2021) One step at a time: the origins of sequential simulation and beyond. Math Geosci 53:193–209. https://doi.org/10.1007/s11004-021-09926-0

Goovaerts P (2005) Geostatistical analysis of disease data: estimation of cancer mortality risk from empirical frequencies using Poisson kriging. Int J Health Geogr 33:4–31. https://doi.org/10.1186/1476-072X-4-31

Goovaerts P (2006) Geostatistical analysis of disease data : accounting for spatial support and population density in the isopleth mapping of cancer mortality risk using area-to-point Poisson kriging. Int J Health Geogr 31:1–31. https://doi.org/10.1186/1476-072X-5-52

Goovaerts P (2008) Accounting for rate instability and spatial patterns in the boundary analysis of cancer mortality maps. Environ Ecol Stat 15:421–446. https://doi.org/10.1007/s10651-007-0064-6

Goovaerts P, Jacquez GM, Greiling D (2005) Exploring scale-dependent correlations between cancer mortality rates using factorial kriging and population-weighted semivariograms. Geogr Anal 37:152–182. https://doi.org/10.1111/j.1538-4632.2005.00634.x

Guillette D, Stratton J, Varia M, Chau V, Loh LC (2020) Canadian public health agency lessons on using choropleth maps to characterize geographic distribution of COVID-19 data. Acta Med Port 33:792–794. https://doi.org/10.20344/AMP.15056

Jaya IGNM, Folmer H (2020) Bayesian spatiotemporal mapping of relative dengue disease risk in Bandung, Indonesia. Springer, Berlin

Journel A, Huijbregts C (1978) Mining geostatistics. Academic Press, New York

Kyriakidis P (2004) A geostatistical framework for area to point spatial interpolation. Geogr Anal 36:259–289

Lucchesi LR, Wikle CK (2017) Visualizing uncertainty in areal data with bivariate choropleth maps, map pixelation and glyph rotation. Stat 6:292–302. https://doi.org/10.1002/sta4.150

Meng Y, Cave M, Zhang C (2019) Comparison of methods for addressing the point-to-area data transformation to make data suitable for environmental, health and socio-economic studies. Sci Total Environ 689:797–807. https://doi.org/10.1016/j.scitotenv.2019.06.452

Pebesma EJ, Wesseling CG (1998) Gstat: a program for geostatistical modelling, prediction and simulation. Comput Geosci 24:17–31. https://doi.org/10.1016/S0098-3004(97)00082-4

R Core Team R Core Team (2021) R: a language and environment for statistical computing

Ribeiro MC, Pereira MJ (2018) Modelling local uncertainty in relations between birth weight and air quality within an urban area: combining geographically weighted regression with geostatistical simulation. Environ Sci Pollut Res 25:25942–25954. https://doi.org/10.1007/s11356-018-2614-x

Rita A, Cristina O, Pereira M, Soares A (2013) Stochastic simulation model for the spatial characterization of lung cancer mortality risk and study of environmental factors. Math Geosci 45:437–452. https://doi.org/10.1007/s11004-013-9443-8

Taylor AR, Watson JA, Buckee CO (2020) Pixelate to communicate: visualising uncertainty in maps of disease risk and other spatial continua, pp 1–6

Waller LA, Gotway CA (2004) Applied spatial statistics for public health data. Wiley, New York

Wickham H (2016) ggplot2. Springer, New York

Young LJ, Gotway CA, Yang J, Kearney G, DuClos C (2009) Linking health and environmental data in geographical analysis: it’s so much more than centroids. Spat Spatiotemporal Epidemiol 1:73–84. https://doi.org/10.1016/j.sste.2009.07.008

Young LJ, Gotway CA (2010) Using geostatistical methods in the analysis of public health data: the final frontier? In: Atkinson PM, Lloyd CD (eds) geoENVVII—geostatistics for environmental applications, quantitative geology and geostatistics, vol 16. Springer, London, pp 89–98

Zaytsev V, Biver P, Wackernagel H, Allard D (2016) Change-of-support models on irregular grids for geostatistical simulation. Math Geosci 48:353–369. https://doi.org/10.1007/s11004-015-9614-x