

W A S A B Y

Water and Soil contamination and Awareness on Breast cancer risk
in Young women

CanMapTool for mapping cancer incidence data

*User manual for cancer
registry's personnel*

Tina Žagar, Korat Sara, Vesna Zadnik,
the WASABY's group of experts on spatial analysis

Version of the CanMapTool: Beta, June 2021

Version of the User manual: 1.1, June 2021

The CanMapTool is developed by Slovenian Cancer Registry in the framework of the WASABY project - *Water And Soil contamination and Awareness on Breast cancer risk in Young women* (<http://www.wasabysite.it/>), which received funding from the 3rd European Union Health Programme under Grant Agreement PP-2-5-2016 (# 769767).

The user manual for CanMapTool called *CanMapTool for mapping cancer incidence data*, whose primarily target audience is cancer registry's personnel, was produced as part of activities in Work package 6 of the WASABY Project and represents Deliverable 6.2.

Development, implementation and user interface of the CanMapTool:

Sara Korat (Slovenian Cancer Registry, Institute of Oncology Ljubljana, Slovenia)

Concept, content and user manual:

Tina Žagar (Slovenian Cancer Registry, Institute of Oncology Ljubljana, Slovenia)

Sara Korat (Slovenian Cancer Registry, Institute of Oncology Ljubljana, Slovenia)

Vesna Zadnik (Slovenian Cancer Registry, Institute of Oncology Ljubljana, Slovenia)

Members of the WASABY's group of experts on spatial analysis:

Tina Žagar (Slovenian Cancer Registry, Institute of Oncology Ljubljana, Slovenia)

Vesna Zadnik (Slovenian Cancer Registry, Institute of Oncology Ljubljana, Slovenia)

Sonja Tomšič (Slovenian Cancer Registry, Institute of Oncology Ljubljana, Slovenia)

Andreja Kukec (Faculty of Medicine, University of Ljubljana, Slovenia)

Sara Korat (Slovenian Cancer Registry, Institute of Oncology Ljubljana, Slovenia)

Ron Pritzkeleit (Institute for Cancer Epidemiology, Universität zu Lubeck, Germany)

Marc Colonna (Iserre Cancer Registry, CHU de Grenoble – Pavillon E, France)

Joséphine Bryere (Université de Caen Normandie, France)

Ludivine Launay (Université de Caen Normandie France)

Fortunato Bianconi (Umbria Cancer Registry, University of Perugia, Italy)

Maurizio Zarcone (Palermo Cancer Registry, Palermo University Hospital, Italy)

Martina Bertoldi (Varese Province Cancer Registry, Fondazione "Istituto Nazionale Tumori", Italy)

Eero Pukkala (Finnish Cancer Registry, Institute for Statistical and Epidemiological Cancer Research, Finland)

Marc Saez (Universität de Girona, Spain)

Citation

Žagar T, Korat S, Zadnik V, WASABY's group of experts on spatial analysis. CanMapTool for mapping cancer incidence data; User manual for cancer registry's personnel. Version 1.1, June 2021. Available online on <http://www.wasabysite.it/>. Accessed on *date*.

This user manual corresponds to CanMapTool version Beta, June 2021

Table of contents

1. Mapping cancer registry's data	4
1.1. <i>Introduction and general objectives of WASABY project</i>	4
1.2. <i>Mapping cancer incidence</i>	5
1.3. <i>Required datasets</i>	5
2. CanMapTool	7
2.1. <i>Installation</i>	7
2.2. <i>Basic information</i>	8
2.3. <i>Population data</i>	9
2.4. <i>Incidence data</i>	11
2.4.1. <i>Geocoding</i>	13
2.5. <i>Deprivation index data</i>	15
2.6. <i>Shapefile</i>	17
2.7. <i>Centroids and grid points</i>	19
2.8. <i>Standardization</i>	21
2.8.1. <i>Age standardized incidence rates (ASR)</i>	23
2.8.2. <i>Standardized incidence ratios (SIR)</i>	23
2.9. <i>Map settings</i>	24
2.9.1. <i>Smoothing</i>	27
2.9.2. <i>Floating weighted averages</i>	27
2.9.3. <i>Bayesian hierarchical modelling</i>	28
2.9.4. <i>Adjusting for covariates</i>	30
2.9.5. <i>Clustering</i>	31
2.10. <i>Output preferences</i>	32
2.11. <i>Creating output</i>	34
3. Disclaimer and limitations of the CanMapTool	37
3.1. <i>Disclaimer</i>	37
3.2. <i>Geocoding</i>	37
3.3. <i>Small areas</i>	38
3.4. <i>Small numbers</i>	38
4. Closing remarks	39
5. Literature	40

1. MAPPING CANCER REGISTRY'S DATA

1.1. INTRODUCTION AND GENERAL OBJECTIVES OF WASABY PROJECT

The use of spatially referenced data in cancer studies is gaining in prominence, fuelled by the development and availability of spatial analytic tools and the broadening recognition of the linkages between geography and health. Understanding spatial patterns of diseases in a population is at the very root of the field of epidemiology. The recent explosion in data gathering, linkage and analysis capabilities fostered by computing technology, particularly geographic information systems (GIS), has greatly improved the ability to measure and assess these patterns. Mapping allows a visualization of areas at high risk that face disparities to help prioritize areas that would benefit from public health actions.

The group of experts on spatial analysis was set up in the framework of WASABY project (<http://www.wasabysite.it/>) to determine the methods suitable for analysis of cancer incidence provided by cancer registries according to availability of data at different spatial aggregation level with emphasis on small-area level and identify open source software(s) applicable in these methods. WASABY project stimulates participating cancer registries to include geocoding into their routine registration process.

The main focus of preparing a common report for participating cancer registries is on visualization and also on promoting reliable maps in both methodological and epidemiological terms to support those cancer registries who do not regularly map their data. Mapping requires some data gathering and data manipulation in many steps and CanMapTool guides through them:

- CanMapTool is free for anyone to download and use. It is written in R Statistical Software version 4.0.3 and the code is accessible to everyone. Because the R versions and packages are constantly developing and so changing, we decided to fix the version of R and used packages, otherwise it would probably stop working some day.
- This user manual for CanMapTool guides the user and is specifically focused on cancer registry's datasets although it can be used much more general.

For this reason, smoothing techniques and adjusting for covariates are also implemented in CanMapTool. There are numerous spatial smoothing techniques – we have selected two very distinctive methods so the differences between the approaches would be most visible and, at the same time, they are visually attractive and regularly applied on cancer registries' data.

1.2. MAPPING CANCER INCIDENCE

Over the last decades, many atlases have been prepared to present cancer burden in specific areas or countries. The progress is supported by increasing availability of GIS and other analytical tools and computer power. Almost all are based on the aggregated data, while the point data are used in few targeted studies for investigating specific questions. The main reason for using the aggregated data is that geocoded data to an x and y coordinates of the residence for both patients and general population are rarely routinely available in the cancer epidemiological research. However, spatially referenced public health data sets have become more available in recent years, especially on the level of small geographical units.

Geographical analyses and mapping differ depending on the available data:

- Area or aggregated data: addresses for cancer cases are aggregated into geographical units, usually administrative areas such as statistical region, municipality, country, postal or zip code. Observations are replaced with group summarization, which can lead to ecological bias and modifiable areal unit problem. The benefit is that no information on the exact address is needed for analyses. In cancer epidemiology choropleth maps are the standard.
- Point data: exact coordinates of residence address (or coordinates of reasonable approximation) for cancer incidence cases and background population (or controls) are required for making inference in cancer epidemiology. Analysis of spatial point patterns are used for dataset with cases and controls. But in case of population data, the cancer cases and population data are not macher (on age and gender) – this is why also geographically changing age structure of population should be considered in the modeling (for example using local SIR estimates or calculating risk surface by generalized additive model).

1.3. REQUIRED DATASETS

Technically, three groups of data are needed in order to use the CanMapTool:

- shapefile: geospatial vector of geographical territory;
- cancer cases: number of cancer cases in specific group, according to age, year and region, but can also be further divided by gender etc.;
- background population: needed for calculating crude rates or producing age standardised maps;
- destination folder must be selected;
- all other datasets and settings are optional.

Additional demands for provided datasets:

- Data on cancer cases, background population and covariates need to be georeferenced to the same geographical level – the same geographical level the provided shapefiles are for.
- Background population data are provided for the same calendar years as cancer cases and for the same geographical units.

- Together with cancer cases and background population cancer registries also need to provide shapefiles for geographical units that are valid for corresponding time period. The shapefile format is a popular geospatial vector data format for GIS softwares and is usually provided by national mapping authorities.

Table of additionally required datasets for each map the user wishes to produce in the output.

- Shapefile, cancer cases and background population are obligatory to upload to the CanMapTool) and are needed for all maps.
- In addition to the listed uploaded data, the user can also upload Centroids and grid points needed for mapping with floating weighted averages method (Finnish map).
- In the table below, all the output figures are listed and minimal requirements are marked.

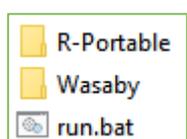
Method and corresponding maps by region	shapefile, cancer cases, background population	data on covariate (EDI)	age standard uploaded by user
Observed incidence <i>Figure 2</i>	✓		
Population <i>Figure 1</i>	✓		
Age-standardized incidence rate (ASR) <i>Figure 3a (EU standard)</i>	✓		
<i>Figure 3b (Segi 1961 standard)</i>	✓		
<i>Figure 3c (standard uploaded by user)</i>			✓
Expected incidence <i>Figure 4</i>	✓		
The European deprivation index (EDI) <i>Figure 8</i>	✓	✓	
The standardized incidence ratio (SIR) <i>Figure 5</i>	✓		
Smoothed SIR without EDI: INLA <i>Figure 9</i>	✓		
Smoothed SIR with EDI: INLA <i>Figure 10</i>	✓	✓	
Smoothed SIR without EDI: Gibbs <i>Figure 11</i>	✓		
Smoothed SIR with EDI: Gibbs <i>Figure 12</i>	✓	✓	
Floating weighted averages (Finnish map) <i>Figure 6 (SIR)</i>	✓		
<i>Figure 7a (ASR, EU standard)</i>	✓		
<i>Figure 7b (Segi 1961, EU standard)</i>	✓		

2. CANMAPTOOL

2.1. INSTALLATION

The CanMapTool is developed in R Statistical Software Version R 4.0.3 (October 2020). For modelling, the following packages were used: igraph version 1.2.6, INLA version 20.12.10, nimble version 0.10.1, sf version 0.9-8. Version 1.5.0 of the Shiny package was used for building an interactive app.

The CanMapTool is open source and freely available to everyone. After the user extracts files from downloaded zipped file, there should be a folder with one file and two folders saved on the computer. All together they take under 4 GB of space on your computer:



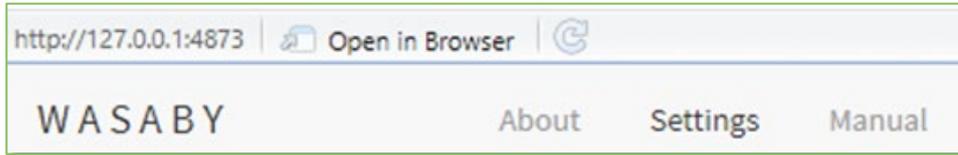
After double clicking the file called 'run.bat':

- At first black CMD window opens.
- The CanMapTool opens in a web browser, which is set to default on the local computer. It takes some time, so please be patient.
- CMD window must be open entire time of running the CanMapTool, which opens in a web browser. In case the user closes the window, the CanMapTool cannot work.
- After the CanMapTool in web browser is closed, also the CMD window can be closed. But in case the CanMapTool needs to be reopened, the CMD window must be closed by user before rerunning the CanMapTool (when it opens again).
- When the user wishes to close the program, he/she must close the opened site in browser and CMD window.

```

C:\Windows\system32\cmd.exe
C:\Users\tzagar\Documents\Register Raka\RR projekti\WASABY\WP6 - CRs\Sarin program\06-21maj2021>set "BINPREF=C:\Users\tzagar\Documents\Register Raka\RR projekti\WASABY\WP6 - CRs\Sarin program\06-21maj2021\R-Portable\rtools40\mingw64\bin\"
C:\Users\tzagar\Documents\Register Raka\RR projekti\WASABY\WP6 - CRs\Sarin program\06-21maj2021>set "BINPREF=C:/Users/tzagar/Documents/Register Raka/RR projekti/WASABY/WP6 - CRs/Sarin program/06-21maj2021/R-Portable/rtools40/mingw64/bin/"
C:\Users\tzagar\Documents\Register Raka\RR projekti\WASABY\WP6 - CRs\Sarin program\06-21maj2021>set "Path=C:\Users\tzagar\Documents\Register Raka\RR projekti\WASABY\WP6 - CRs\Sarin program\06-21maj2021\R-Portable\rtools40\mingw64\bin;C:\Users\tzagar\Documents\Register Raka\RR projekti\WASABY\WP6 - CRs\Sarin program\06-21maj2021\R-Portable\rtools40\usr\bin;C:\Users\tzagar\Documents\Register Raka\RR projekti\WASABY\WP6 - CRs\Sarin program\06-21maj2021\R-Portable\bin;C:\Program Files (x86)\Common Files\Oracle\Java\javapath;C:\Windows\system32;C:\Windows;C:\Windows\System32\wbem;C:\Windows\System32\WindowsPowerShell\v1.0;C:\Windows\System32\OpenSSH;C:\Program Files\Gemalto\Classic Client\BIN;C:\Program Files (x86)\Gemalto\Classic Client\BIN;C:\Program Files\IBM\SPSS\Statistics\24\JRE\bin;C:\Users\tzagar\AppData\Local\Microsoft\WindowsApps;"
C:\Users\tzagar\Documents\Register Raka\RR projekti\WASABY\WP6 - CRs\Sarin program\06-21maj2021>"C:\Users\tzagar\Documents\Register Raka\RR projekti\WASABY\WP6 - CRs\Sarin program\06-21maj2021\R-Portable\R-4.0.3\bin\R.exe" CMD BATCH "Wasaby\StartShiny.R"
  
```

When CanMapTool opens in web browser, three options are available on the menu:



- About : gives some general information about CanMapTool.
- Settings : through all the given options the user sets text describing the dataset, imports the datasets, and selects desired settings for the output.
- Manual : it enables downloading this document with detailed instructions.

The 'Settings' are divided into **nine sections**.

- The sections are organized in such a way, that the app guides the user what to do. Each section also has an icon 'i', which opens a window with some information on the section.
- Each section has 'Edit' button for modifying the settings and uploading the datasets in a current section only. The changes are submitted with 'Save' button or dismissed with 'Cancel' button.
- Once one of the section is in 'Edit' mode, the others can not be edited. In case the user forgets to 'Save' the current section or cancel editing, the warning appears:



- Each section has a detailed description in the following subchapters of this 'Manual'.

At the bottom of the 'Settings' page, there is a 'Create output' button. After all settings on the page are edited, the user uses this button to start modelling the selected analyses all at once. Details are described under the chapter 2.11 Creating output.

2.2. BASIC INFORMATION

In the first section of the CanMapTool the user writes general information about the dataset and using the CanMapTool, which is then used in automatically generated description written at the beginning of the Word document, which is an output of the CanMapTool:

- Title: is used as a title in the Word document. It is also used for naming the output files. We recommend this text is as short as possible.
- Author: is used in automatically generated description text.

- Institution: is used in automatically generated description text.
- Date: is set automatically to current date, but the user can change this date it.
- Description: is used as a subtitle. This can be longer text, more than 5,000 are accepted.

For example: this input of the basic parameters:

Basic information ⓘ

Cancel
Save

Title:

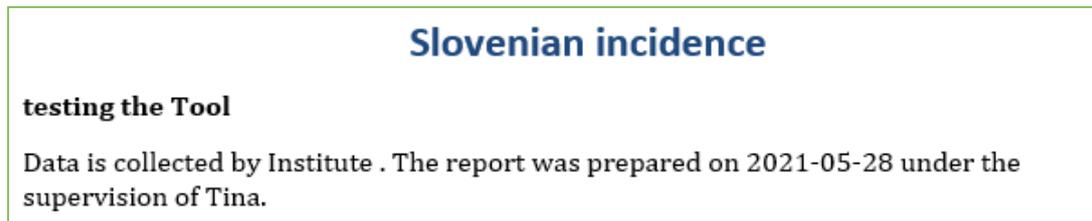
Author:

Institution:

Date:

Description:

... gives such appearance of the Word document, which is an output of the CanMapTool:



2.3. POPULATION DATA

The population data contains information on the background population from which you collect the data on cancer patients – meaning this is the population covered by the cancer registry.

- Country or regional statistical office (or other public authority) usually provides population numbers.
- Population for each region (or area used in the analysis) is required.
- Population for each age or 5-year age groups in every region is required.
- Population for a whole observed period is required for each separate calendar year for which incidence data is provided in every region is required. Information on the calendar year needs to be in a separate column. If population data is not provided for the entire observed period, it is estimated from the data for the provided year(s). The input of population data for at least one calendar year is required.

- In case the user wants to perform analysis separately by gender, this information needs to be included in separate columns in the population file (and not as an additional variable on gender). In this case, the user needs to rerun the modelling separately for each gender, each time selecting appropriate columns in population and incidence sections.
- The population data is the most important input data right after the incidence data. Without the population data, CanMapTool cannot create an output.
- In Figure 1 of the CanMapTool output, the population data by region is mapped for the purpose of inspection. If population data for more than one calendar year is provided, for each region the average population is mapped.

In this section the population data should be uploaded:

- The user selects the data file from the computer.



- Excel and csv files are accepted.
- Excel files are preferred by CanMapTool. Csv files are also accepted but the user must ensure that for the field separator comma is used and for numbers, decimal point is used.

The necessary columns are:

- Age: can be provided in years (completed years of age without decimal places) or five-year age groups starting from age 0 in the following notation: 0-4, 5-9, 10-14, 15-19, ..., 75-79, 80+.
- Region ID or Region: It is important that ID or names for regions are unique otherwise CanMapTool produces an error and maps are not produced. Also names of the regions are accepted as input (type of data is 'Region'). Any kind of string is accepted, but the CanMapTool works only if the IDs (or names) are exactly matching the IDs in all other files (for incidence, covariate and shapefile).
- Pop_number: this is aggregated number for population counts in a given year, age (or age group) and region ID.
- Year: years for the population data, which should match cancer incidence data file.
- Any additional columns are ignored.

Uploading the population file:

- After the data is loaded, it is necessary to **select the appropriate column type** (first row) and the corresponding column names (second row) in the original data file. Types of data are set to default in the first row. In a second row, the first column from your data is automatically chosen.

Select your type of data:

AGE_GROUP ▼	YEAR ▼	REGION_ID ▼	POP_NUMBER ▼
-------------	--------	-------------	--------------

Select columns from your data that match the data types above (in the right order):

YEAR ▼	YEAR ▼	YEAR ▼	YEAR ▼
--------	--------	--------	--------

- After the user clicks save, a **data preview** is displayed.

Show entries Search:

AGE_GROUP	REGION_ID	POP_NUMBER	YEAR
0-4	3000101001	21	1996
0-4	3000101002	32	1996
0-4	3000101003	0	1996
0-4	3000102001	12	1996
0-4	3000102002	8	1996
0-4	3000102003	10	1996

Showing 1 to 6 of 207,400 entries

PREVIOUS 1 2 3 4 5 ...

34567 NEXT

2.4. INCIDENCE DATA

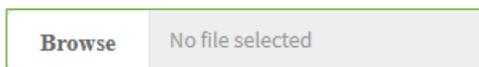
Incidence is the absolute number of all newly diagnosed cases of any disease in a defined population in one calendar year. The incidence considers the number of cases of a disease, not the number of patients, therefore the same patient may contribute more than one disease case to the incidence number if a person is diagnosed with more than one different cancer.

- Data on cancer incidence is usually provided by cancer registries.
- Incidence for each region (or area used in the analysis) is required.
- Incidence for each age or 5-year age groups in every region is required.

- Incidence is required for each separate calendar year. Information on the calendar year needs to be in a separate column. The input of incidence data for at least one calendar year is required.
- In case the user wants to perform analysis separately by gender, this information needs to be included in separate columns in the incidence file (and not as an additional variable on gender). In this case, the user needs to rerun the modelling separately for each gender, each time selecting appropriate columns in population and incidence sections.
- The incidence data is one of the most important input data. Without the incidence data CanMapTool produces an error.
- In Figure 2 of the CanMapTool output the incidence data by region is mapped only for the purpose of inspection. If incidence data for more than one calendar year is provided, for each region the sum of incidence in specific region is mapped in Figure 2. Please see also the chapter 3 'Disclaimer and limitations of the CanMapTool' for issues on data protection.

In this section the population data should be uploaded:

- User selects the data file from the computer.



- Excel and csv files are accepted.
- Excel files are preferred by CanMapTool. Csv files are also accepted but the user must ensure that for the field separator comma is used and for numbers, decimal point is used.

The **necessary columns** are:

- Age: can be provided in years (completed years of age without decimal places) or five-year age groups starting from age 0 in the following notation: 0-4, 5-9, 10-14, 15-19, ..., 75-79, 80+ .
- Region ID: It is important that ID or names for regions are unique otherwise CanMapTool produces an error and maps are not produced. Also names of the regions are accepted as input (type of data is 'Region'). Any kind of string is accepted, but the CanMapTool works only if the IDs (or names) are exactly matching the IDs in all other files (for incidence, covariate and shapefile).
- Cases: this is aggregated number for counts of cancer cases (the incidence) in a given year, age (or age group) and region ID.
- Year: years for the incidence data, which should match population data file.
- Any additional columns are ignored.

Uploading the incidence file:

- After the data is loaded, it is necessary to select the appropriate column type and the corresponding column names in the original data file. By default, the first type of data (first row) and the first column (second row) is selected for all four required types of data.

Select your type of data:

AGE_GROUP YEAR REGION_ID CASES

Select columns from your data that match the data types above (in the right order):

Age_class Year CUSEC Number

- After the user clicks save, a data preview is displayed.

Show entries Search:

YEAR	AGE_GROUP	REGION_ID	CASES
1996	25-29	3002701025	1
1996	25-29	3003004007	1
1996	25-29	3003005036	1
1996	30-34	3001201004	1
1996	30-34	3001401001	1
1996	30-34	3001602012	1

YEAR AGE_GROUP REGION_ID CASES

Showing 1 to 6 of 2,329 entries PREVIOUS 1 2 3 ... 389 NEXT

2.4.1. Geocoding

The incidence registered by cancer registries only includes the data on patients with permanent residence in the registry's area at the time of diagnosis (regardless of the place where they have been treated). Case-specific data relevant for geographical analysis are: information on cancer case identifier, age at diagnosis, gender and address. The address per se is not included in geographical analyses of cancer incidence as it is usually recorded as an alphanumeric variable. It serves as information for further categorization of cancer case into one of the sub-regions covered by cancer registration.

Cancer registries follow a common set of rules and recommendations for cancer registration and coding information on cancer entities. However, information on the address is not unified across

cancer registries, mainly because of different situations specific for each country, which influences the accessibility of detailed address. Addresses can be available in different forms (for example coded or only alphanumerically transcript) and levels (for example only postal code or detailed information on address including coordinates of the dwelling) – both have an impact on the level of geographical regions the analyses are possible on (larger regions, municipalities, postal codes, etc.) and further on the possible analyses selection (using large vs. small geographical areas).

Geocoding is the process of converting addresses (e.g., a street address) into geographic coordinates (e.g., geographical latitude and longitude) which one may use to place markers on a map, or to position a map. For example, the Cancer Centre for Normandy – Centre François Baclesse - located in 3 Av Général Harris, 14076, Caen, France - has geographical coordinates (49.203529, -0.354513) in WGS84 coordinate system. Having precise coordinates for patients included in a study allows to have information for all geographical units: from the smallest to the largest ones (IRIS for France). It is necessary to evaluate accessibility to health care centres or health professionals (for example using distance), to perform environmental studies which aim to determine the effect of given pollutants on health and to calculate ecological deprivation indexes. In previous example, geographical coordinates correspond to IRIS number 141181404. Geocoded addresses can be subsequently aggregated to any larger geographical units (administrative or user defined), which is beneficial when dataset cover longer time periods that are subject to changes in administrative geographical units.

Preparation for geocoding:

- The required information is a precise address – registered by cancer registries – including house number, street type, street name, postal code, municipality or city (other code specific to country, for example in France, insee code which is specific for each municipality). Be aware that this information is considered as directly identifying an individual by national data protection authority, so specific authorization might be required. The information may either be available in a common field or in separate fields, according to software.
- Geographic Information System (GIS). The most famous commercial GIS are Mapinfo® (Pitney bowes) and ArcGIS® (ESRI). QGIS is a freely available GIS and is more and more used by researchers (<https://qgis.org/en/site/forusers/download.html>). Other programmes include qVSI or GRASS GIS (the list is not exhaustive). As QGIS is widely used, there are many tutorials and discussions on the internet. This should be the most appropriate for discovering GIS.Maps.
- To make the link between address and geographical coordinates one needs maps supported by GIS. Some are commercialized by famously established editors (for example ESRI) but the price can be quite high if you do not use it regularly. Free data are available on OpenStreetMap (www.openstreetmap.org).

The geocoding process is performed step by step. Address is used with all information (number, type, street name, postal code, locality). If an exact match is found, the software goes to the next

address. If not, the next location level (using only type, street name, postal code and locality) is considered, and so on. Geolocation may be done at different levels according to the information used or available:

- Level 1: Number, type, street name, postal code, locality;
- Level 2: Type and street name, postal code, locality;
- Level 3: Postal code, locality (municipality level, often the city hall).

2.5. DEPRIVATION INDEX DATA

Case-specific data relevant for geographical analysis are: information on cancer case identifier, age at diagnosis, gender and address. Other data describing cancer entity, personal characteristics or environment may be additionally included in analyses as covariates. At the current development of CanMapTool, only one covariate can be included into the analyses and it has to be in quintiles or continuous variable, which is transformed into quintiles by CanMapTool.

An unequal distribution of well-known major risk factors in population typically explains much of the variation in the cancer incidence worldwide. Socioeconomic deprivation is recognised as one of the important predictors of cancer risk. Estimating the burden of cancers attributable to socioeconomic inequalities enables us to evaluate the gains that could be achieved by implementing targeted public health interventions.

The deprived are those who lack the necessities and activities that are widely encouraged or approved in the society to which they belong. Such unmet needs are due to a lack of resources of all kinds, not just financial. Needs differ between societies and periods. By following the Townsend philosophy of relative deprivation and its extension to population level on an ecological scale, a European Deprivation Index (EDI) was proposed by two French teams in 2012. They suggested a method for constructing a country-specific ecological deprivation index that best reflects individual experience of deprivation by using the European Union Statistics on Income and Living Conditions survey (EU-SILC) and selects ecological variables from national censuses that are the most closely related to the individual deprivation indicator specific for each country. The procedure can be used to construct an ecological deprivation index using the smallest available geographical levels in a replicable way for all European Union members and has since then been used in several studies on social inequalities in cancer burden, screening uptake and health care access, orthopaedic care and even environmental exposure. EDI is not provided by authorities but is calculated by EDI has been developed for many additional European countries also in the framework of the WASABY project. These are the reasons EDI is included into the CanMapTool as primary covariate.

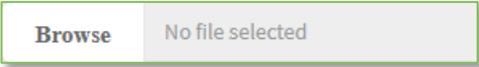
EDI is usually calculated for one reference year, which may correspond to census performed in individual country. For this reason, only one set of EDI per regions is required by CanMapTool and

not for each incidence year (as it is required for population data). From quantitative form of the EDI for each region, the categories for EDI are calculated as quintiles:

- 1 - low deprivation
- 2 - medium-low deprivation
- 3 - medium deprivation
- 4 - medium-high deprivation
- 5 - high deprivation

In this section the EDI data should be uploaded:

- User selects the data file from the computer.



- Excel and csv files are accepted.
- Excel files are preferred by CanMapTool. Csv files are also accepted but the user must ensure for the field separator comma is used and for numbers, decimal point is used.

The necessary columns are:

- Region ID: It is important that ID or names for regions are unique otherwise, CanMapTool produces an error and maps are not prepared. The format of Region/Region ID has to correspond to the format of Region/Region ID in submitted shapefile.
- Deprivation_index or deprivation_index_category: In case the numerical EDI is provided (EDI in its quantitative form), CanMapTool calculates quintiles. This is the same as in case the user provides categories representing the quintiles of the EDI (as column deprivation_index_category) but the values must be numeric (accepted values for categories are 1, 2, 3, 4 and 5).
- If the uploaded data file has some additional column (for example labels of the EDI quintiles), CanMapTool is not disturbed by it and any additional columns are ignored.

Uploading the EDI file:

- After the data is loaded, it is necessary to select the appropriate column type and the corresponding column names in the original data file. By default, the first type of data (upper row) and the first column (second row) is selected for both required types of data.



- After the user clicks save, a data preview is displayed.

Show Search:

REGION_ID	DEPRIVATION_INDEX	DEPRIVATION_INDEX_CATEGORY
3000101001	3.79809233922795	5
3000101002	-0.833553081748936	3
3000102001	1.20738384338046	4
3000102002	0.210763758307195	3
3000102003	-1.66738366517041	2
3000201001	2.45325775102682	5

Showing 1 to 6 of 861 entries

2.6. SHAPEFILE

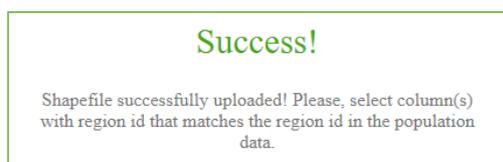
The shapefile format is a geospatial vector data format for geographic information system (GIS) software. It is developed and regulated by Esri as a mostly open specification for data interoperability among Esri and other GIS software products. The shapefile format is a digital vector storage format for storing geometric location and associated attribute information. This format lacks the capacity to store topological information. The shapefile format was introduced with ArcView GIS version 2 in the early 1990s. It is now possible to read and write geographical datasets using the shapefile format with a wide variety of software. The shapefile format stores the data as primitive geometric shapes like points, lines, and polygons. These shapes, together with data attributes that are linked to each shape, create the representation of the geographic data. The term "shapefile" is quite common, but the format consists of a collection of files with a common filename prefix, stored in the same directory. The three mandatory files have filename extensions .shp, .shx, and .dbf. The actual shapefile relates specifically to the .shp file, but alone is incomplete for distribution as the other supporting files are required. (source: <https://en.wikipedia.org/wiki/Shapefile>).

First step:

- The user shows the folder with a shapefile to be uploaded into CanMapTool. Shapefile is usually provided by country or regional mapping authority.
- **IMPORTANT:** it is necessary to select a folder that is identically named as the included three mandatory files (shapefile format).



- The three mandatory files (shapefile format) have filename extensions (source: <https://en.wikipedia.org/wiki/Shapefile>):
 - *.shp — shape format. Includes the feature geometry itself. It is recommended that the user does not change the content.
 - *.shx — shape index format. Includes a positional index of the feature geometry to allow seeking forwards and backwards quickly. It is recommended that the user does not change the content.
 - *.dbf — attribute format. The content of this file is more familiar to the user since it resembles Excel and csv format. It contains columnar attributes for each shape, in dBase IV format. With certain knowledge, the user can change the content, but it is recommended only to add columns to existing ones with additional data (for example area of shape, population in each shape, etc.). It is important the user do not change ID of shapes or change the sorting.
- Two notices are produced to inform the user about the progress and some of the steps takes longer time.
- Uploading shapefile might take a while. After selecting the folder that contains a shapefile, wait for the notification that the upload has finished.

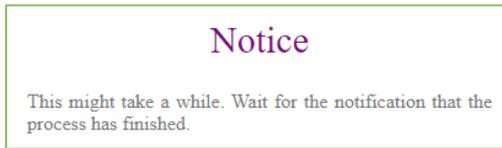


Second step:

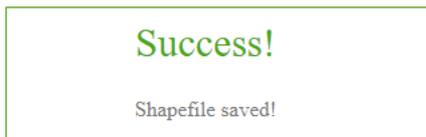
- In CanMapTool user must select the name(s) of the columns with the region codes, which must match the codes from the other data sources (population and incidence data).
- **IMPORTANT:** the names of the regions must be uniquely defined and repeated only once, otherwise CanMapTool can not work!

Third step:

- After clicking on 'Save', a notification is displayed that it is necessary to wait for the shapefile to load.



- When loaded, a notification appears stating that a shapefile has been successfully loaded.

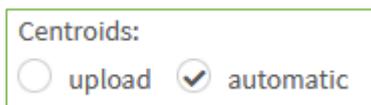


2.7. CENTROIDS AND GRID POINTS

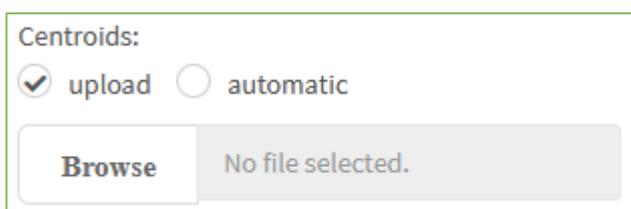
Centroids and grid points are needed only for preparing maps by floating weighted averages method (Finnish maps). It is recommended that CanMapTool calculates both as the function is so optimized that calculations take less than a minute. In case these calculations would take too long (for example with large number of regions in analysed area), the user also has an option to upload these data.

Centroids:

- Centroids are X and Y coordinates of centroids for regions provided by user in the shapefile.
- Centroids are calculated by CanMapTool if user selects option 'automatic'. This is default option and user does not need to do any settings in this section. This option is recommended.



- User can also upload centroids for regions as an Excel file. Three columns are required: ID for regions, X and Y coordinates. It is important that ID for regions are unique and exactly match IDs in shapefile, otherwise CanMapTool produces an error and maps are not produced.



- In case maps by floating weighted averages method are selected and successfully modelled (meaning output maps are prepared), a preview is also displayed (but only after refreshing of CanMapTool, i.e. after reopening it). Number of entries equals the number of the regions in the shapefile.

Show <input type="text" value=""/> entries		Search: <input type="text" value=""/>
X	Y	REGION_ID
632227.8	4211136	3002901012
632239.9	4210760	3002901013
633183.3	4211227	3002901014
626577.9	4216339	3002901015
662806.9	4207505	3003001001
663210.4	4207283	3003001002

X Y REGION_ID

Showing 1 to 6 of 1,220 entries

PREVIOUS 1 2 3 ... 204 NEXT

Grid points:

- Grid points are X and Y coordinates for equally spaced square grid covering entire area provided by user in the shapefile. Grid does not depend on regions.
- Number of grid points influences the granulation of map prepared by floating weighted averages method. On map grid point are presented as coloured pixels.
- Grid points are calculated by CanMapTool if user selects option 'automatic'. This is default option and the user does not need to do any settings in this section. This option is recommended.

Grid points:

upload automatic

- In automatic option approximately 25,000 grid points are generated. This means, that the spacing between grid points depends on how large area is covered in uploaded shapefile. The implemented procedure first produces a rectangle around the entire area and generates 30,000 grid points. In second step points outside the area are discarded, So number of grid points depends on the shape of an area (the proportion of the area in the rectangle).

- User can also upload coordinates for grid points as an Excel file. It is important that grid covers area in the provided shapefile – CanMapTool does not verify the appropriateness of these coordinates.

Grid points:

upload automatic

Browse No file selected.

- In case maps by floating weighted averages method are selected and successfully modelled (meaning output maps are prepared), a preview is also displayed (but only after refreshing of CanMapTool, i.e. after reopening it). Number of entries equals the number of grid points and depends on how big is the area.

Show entries Search:

x	y
632091.83177328	4210795.97540353
619087.941620964	4207792.85561155
619521.404626041	4208543.63555955
619954.867631118	4209294.41550754
619954.867631118	4222808.45457145
619954.867631118	4224310.01446745

Showing 1 to 6 of 17,373 entries

PREVIOUS
1
2
3
...
2896
NEXT

2.8. STANDARDIZATION

When standardizing by age in epidemiological analyses we use one of the internationally agreed standard or standard specific for a country or study. CanMapTool enables to use three options for age standardization:

- World standard population (Segi 1961),
- European standard population (Doll 1976),
- Standard uploaded by user – it is important that values are given for all age groups otherwise CanMapTool produces an error.

European Standard

Segi 1961 Standard

Upload Standard

Browse

No file selected

In the table below World (Segi 1961) and European (Doll 1976) standard population are given.

Age group	World standard (Segi 1961)	European standard (Doll 1976)
0-4 years	12,000	8,000
5-9 years	10,000	7,000
10-14 years	9,000	7,000
15-19 years	9,000	7,000
20-24 years	8,000	7,000
25-29 years	8,000	7,000
30-34 years	6,000	7,000
35-39 years	6,000	7,000
40-44 years	6,000	7,000
45-49 years	6,000	7,000
50-54 years	5,000	7,000
55-59 years	4,000	6,000
60-64 years	4,000	5,000
65-69 years	3,000	4,000
70-74 years	2,000	3,000
75-79 years	1,000	2,000
80-84 years	500	1,000
85+ years	500	1,000

2.8.1. Age standardized incidence rates (ASR)

Age standardization is used in epidemiological analyses since it considers not only the distribution of population but also their age structure. Age-standardized rate (ASR) is a method of direct standardization that takes into account the period of diagnosis and age structure of population. ASR is a theoretical incidence rate assuming that the age structure in the observed population is the same as in the standard population – it tells the crude rate in observed population in case if it's age structure is the same as in standard population. Age-standardized rate is used when analysing the incidence/mortality within a longer time of period (if the age structure of population changes in time) or comparing the incidence/mortality between populations with different age structure.

Crude cancer incidence rate (crude rate) is defined as number of new cases in a specified time period, divided by the number of persons, living in observed area specific population in the same time interval and geographical unit. Age-specific rate is a crude rate for specific age groups. Age-standardised rates (ASR_i for geographical unit i) are calculated by multiplying the age-specific rates by standard population weights and then summing together. Crude rates and ASRs are usually expressed as the number of cancers per 100,000 population at risk.

For the most part, the choice of weights (the standard population) is based on convention, the intended and potential comparisons, and various other considerations. There is often no absolute correct choice, and there can easily be different opinions about the best one. Regardless of the chosen standard population, the ASRs do not reflect the true cancer burden on the population but serves as relative estimation of the magnitude of cancer burden for the purpose of comparisons.

2.8.2. Standardized incidence ratios (SIR)

We assume the observed number of new cancers in each single geographical unit follows a Poisson distribution. Expected number of new cases are calculated as if the population in a particular area has the same age-specific incidence rates as some larger comparison population, usually the overall population of the whole study area, or some other reference population. Expected number is derived from indirect standardization. Observed and expected numbers of cases can be compared, because both refer to the same population. The ratio of the observed number of cases to that expected is called standardized incidence ratio (SIR).

SIR of 1 indicates that the total observed number of cases is the same as expected in the geographical unit being studied compared to age-specific rates in reference population. A ratio less than 1 indicates a lower than average relative risk and over 1 is a higher than average. This also means SIR maps can not be compared among themselves except in case they are all produced with same reference age-specific rates – for example, in time trends reference age-specific rates can be

taken for whole time interval under study, but maps are prepared for individual shorter time periods.

As with the direct method, the result depends in part upon the chosen standard (reference population). However, the indirect method of standardization is less sensitive to the choice of standard than the direct one. Indirect method is also preferable to the direct method when age-specific rates in geographical unit is based on small numbers of subjects – rates used in direct adjustment would thus be open to substantial sampling variation.

2.9. MAP SETTINGS

First, the user selects the desired maps:

- Selection of maps affects the modelling performed by CanMapTool (for example, if the Finnish map is not selected, no smoothing will be performed within the program and the computation will be faster).
- In case information on EDI is not provided, the user can deselect all relevant maps where EDI is required. But this is not necessary (maps are not produced if EDI is not provided, CanMapTool does not report errors).

<input checked="" type="checkbox"/>	Observed incidence by region
<input checked="" type="checkbox"/>	Population by region
<input checked="" type="checkbox"/>	Age-standardized incidence rate by region
<input checked="" type="checkbox"/>	Expected incidence by region
<input type="checkbox"/>	The deprivation index EDI
<input checked="" type="checkbox"/>	The standardized incidence ratio (SIR) by region
<input checked="" type="checkbox"/>	Smoothed SIR without EDI: INLA by region
<input type="checkbox"/>	Smoothed SIR with EDI: INLA by region
<input checked="" type="checkbox"/>	Smoothed SIR without EDI: Gibbs by region
<input type="checkbox"/>	Smoothed SIR with EDI: Gibbs by region
<input type="checkbox"/>	Finnish map(s)

By clicking the button '**More settings**', additional options appear. These additional options can be hidden again by clicking the button '**Less settings**'.



CanMapTool is general tool and the user can use it for whichever level of areal division at his/her wish. When CanMapTool is used on small regions in (areal) size the colouring might not be visible.

- For this purpose the 'Map settings' section offers mapping with hidden region lines (which are always in grey colour) by selecting this option:

hide region lines on the output maps

- The other possibilities are, that the user either provides the selection of the dataset and corresponding regions to the CanMapTool or
- Use the numerical output in Excel (in 'Output preferences' section there is an 'export .xlsx file with data' option for this purpose) and map the required selection of regions by themselves.

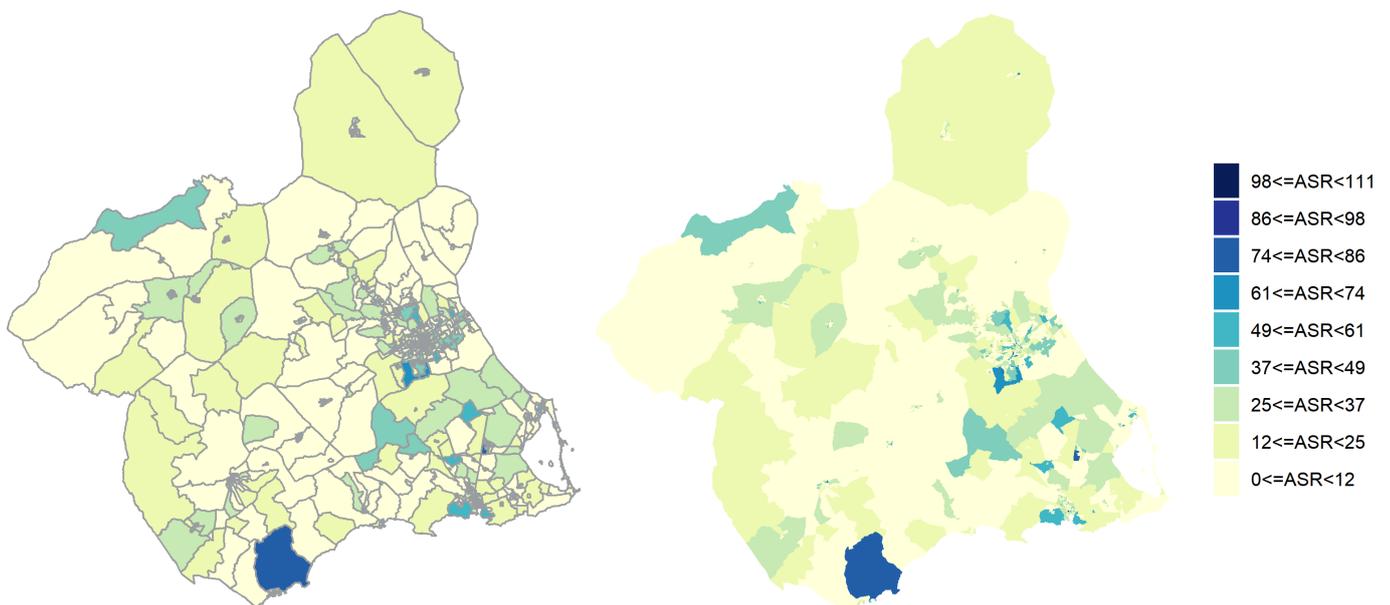


Figure: Age-standardized (Segi 1961 standard) incidence rate (ASR) per 100,000 population by region, Murcia 1996-2021 breast cancer in females with age at diagnosis from 15 to 49 years. Maps are the same but one is without lined borders (on right) and one is mapped with borders in gray colour (on left).

With 'More settings', the desired colour palette can be selected for those maps that are selected. For ASR and SIR (where there are multiple maps), it is possible to keep the same colour palettes for all SIR or ASR maps. In case the option 'keep the same settings for all ASR maps' is not selected, the options appears for all ASR maps selected by user.

Mapping ASR:

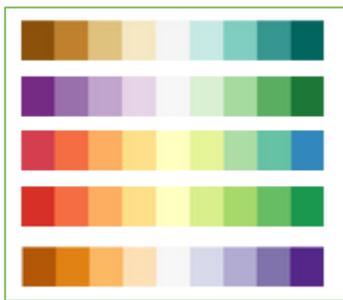
- For mapping ASR five sequential palettes are available:



- User can choose the number of categories from 3 to 9.
- For each map the scale is calculated separately. The boundaries are calculated so, the intervals are even depending on the number of categories (user can not change them).
- Same colour palettes are available for continuous presentation of ASR on maps.

Mapping SIR:

- For mapping SIR five diverging palettes are available:



- For mapping SIR the boundaries are fixed and user can not change them.

For mapping floating weighted averages (Finnish maps) some additional settings are available:

- Parameter for influence of the neighbourhood has three levels (more, medium and less intense smoothing). This parameter is the distance at which weighting function drops by half. If this parameter is set to less intense smoothing, the single geographical unit's specific rates (also geographical units with small number of cases) become more pronounced.
- Parameter for maximum influence has three levels (more, medium and less intense smoothing). This parameter controls the maximum distance set in distance weighting function (i.e. up to which distance the regions are included in calculation of the parameter for a specific grid point). Changing this parameter usually does not affect results much, since the spatial weights are diminishing fast by distance.
- Number of excluded regions can be set from zero to ten.
 - Default number is set to 5.
 - This is the number of big regions excluded from smoothing

- These big regions are instead presented on the map as a circles (presenting the region instead of its shape) positioned at the centroids of the regions.
- The colour of the circles corresponds to the input value and not smoothed value.
- The size of the circle is proportional to the population number in the specific region (it is calculated as four-times population number divided by minimum population among the excluded regions) and can not be changed by the user.

2.9.1. Smoothing

Geographical units are problematic in terms their size and the population they cover is very different over the area. If large spatial units are used, the heterogeneity of exposure and different population characteristics may be missed. On the other hand, the number of cancer cases is usually low in small spatial units and analysing the observed spatial pattern proves to be inefficient, as the population base from which these cases arise is often very low too. This can lead to unstable and misleading estimates of the true rate. Modern approaches to relative risk estimation often rely on smoothing methods, which produce more stable and “less noisy” estimates, providing more confidence that any observed differences are real and not just due to chance.

The basic idea of mapping the smoothed ratios is to borrow information from neighbouring regions to produce more stable estimate of the ratio associated with each geographical unit and thus separate out the spatial pattern from the noise. Smoothing techniques are appropriate when we are not looking for individual regions with elevated ratios but, instead, we are interested in getting the general assessment of broad trends and patterns. On the other hand, smoothing might remove details from the map that would be important for interpretation. If the data reflect region specific features (when cancer risk determinants depend on local administrative decisions), smoothing is not advisable.

There are numerous spatial smoothing techniques – we selected three very distinctive methods so that the differences between the approaches would be most visible and, at the same time, they are visually attractive and regularly applied on cancer registries’ data.

2.9.2. Floating weighted averages

The “Finnish smoothing method” uses floating weighted averages and was first used in the national cancer incidence atlas of Norway and further developed in the Finnish atlas and in the Cancer Atlas of Northern Europe. The floating weighted averages method has mostly been applied to age-adjusted incidence and mortality rates (direct standardization) but can equally well be used for many other measures of cancer frequency such as to SIR. Floating weighted averages aim at

diminishing the random variation by locally calculating floating averages, weighted by population and by distance.

Because of population weighting, large cities (that is, municipalities which include large cities) significantly influence its neighbourhood when smoothing. It has been shown that cancer burden can vary between the main cities and the surrounding less urbanised regions. For this reason, selected big cities are often excluded from the smoothing and illustrated separately on the map as circles whose colour presents observed (i.e. non-smoothed) cancer incidence and the area corresponds to the population size in those cities. The procedure thus minimizes their strong effect in the bias of the estimates in their surroundings. In addition, the excluded big cities are preferable to be positioned at the centroids of the principal (or the biggest) cities themselves rather than at the centroids of the corresponding geographical units. Further adjustment for more relevant cancer maps is to position all geographical units to coordinates (centroids) of principal city/settlement instead of centroid of the geographical unit itself, which better accounts for the population distribution in floating weighted averages method.

2.9.3. Bayesian hierarchical modelling

Another widely used approach to handle unreliable observations in the spatial analyses is the Bayesian hierarchical modelling. There are numerous ways to conduct spatial smoothing within Bayesian models, including through considering distance between areas, or adjacency. The general concept used in the models involves defining a neighbourhood of adjacent areas for each of the small areas, such that the estimate for a given area is dependent on the areas it shares a boundary with, making the estimate more similar to those of its neighbours. Areas which have small populations will be subjected to greater neighbourhood smoothing compared to areas with larger populations.

Prior distributions are assigned to random effects and hyperprior distributions are assigned to the parameters of the prior distributions, thus creating a multilevel hierarchical Bayesian model. The posterior distribution is the target outcome and is approximately equal to the prior times the likelihood.

BYM Bayesian hierarchical modelling

In the classic model of Besag, York and Mollié, BYM with Markov Chain Monte Carlo (MCMC) methods, spatially structured variation is not independent of unstructured variation (a problem called non-identifiability). As a consequence, part of the spatial dependence (structured variation) might result as quite heterogeneous (unstructured variation) and vice versa. The convolution model

originally proposed by Besag et al.:

$$O_i \sim \text{Poisson}(\mu_i)$$

$$\ln \frac{O_i}{E_i} = \ln \mu_i = \ln E_i + a + H_i + S_i$$

where a represents the basic (logarithmic) relative risk of disease in the entire study area. O_i and E_i represent the observed and the expected number of cases in the i -th geographical unit. H_i and S_i are two types of random effects, which handle the variation that cannot be explained by fixed effects. H_i represents the unstructured component that is geographically independent. H_i is given the independent normal distribution with mean zero and precision τ_h . The spatial autocorrelation component (S_i) is defined according to the conditional autoregressive (CAR) model of Besag, York, and Mollie, where τ_s controls smoothing induced by this prior (larger values smoothing more than smaller ones). The heterogeneous component H was assigned a normal probability distribution with average zero and an accuracy τ_h . In the Poisson count case the commonest assumed prior distribution is that precision parameters τ_s and τ_h have Gamma priors (0.5,0.0005) as suggested by Bernardinelli et al. The choice of gamma hyperprior probability distribution parameters did not affect the final results.

In case there are less than 2,000 regions the number of iterations drops to 1,000 iterations and the first 600 iterations were discarded as “burn-in” samples. This reduction is for the purpose of speed up the calculations and does not effect on the results. The user can not control these two parameters.

INLA and BYM2

There are alternative formulations to the BYM model, such as the Leroux and Dean models, in which it is ensured that the structured spatial variation is independent of the unstructured. However, neither model scales spatial variation. As a consequence, hyper parameters depend on the spatial structure of the problem and cannot be interpreted correctly. On the other hand, inferences will be made using a Bayesian approach. In this context, the choice of a priori distributions of hyper parameters, known as priors, can have a considerable impact on the results. Leroux and Dean models use standard priors that lead to overfitting. The main consequence of overfitting (a problem also known as multicollinearity in the context of multiple linear regression) is that the estimators of the variances are greater than the real ones and, therefore, the credibility intervals will be much wider than expected, which implies that the null hypothesis (that the coefficients are equal to zero) will not be rejected more times than it should.

Simpson et al. proposed a modification of the BYM model (BYM2) that solves these problems, because it scales spatially structured variation and uses priors that penalize complexity (called PC priors). These priors are robust, in the sense that they have no impact on the results and also have an epidemiological interpretation.

MCMC is slow (often very slow), it does not scale well, and it sometimes fails with complex models (model will not converge). In this sense, Integrated Nested Laplace Approximations (INLA) is a (very) fast alternative to MCMC for the general class of latent Gaussian models. In addition, the use of PC-priors (in INLA) allows the results not to depend on the priors (as does the MCMC). The Integrated Nested Laplace Approximations (INLA) approach is implemented in the R package R-INLA. The fundamental building block of such Gaussian Markov random field (GMRF) models, as implemented in R-INLA, is a high-dimensional basis representation, with simple local basis functions.

2.9.4. Adjusting for covariates

Ecological analysis is defined as the assessment of the associations between disease incidence (eg, suicide) and variables of interest (eg, social or environmental covariates). These variables in an ecological analysis are defined on aggregated groups of individuals rather than the individuals themselves. The reason for focusing on the comparison of groups rather than individuals is that individual-level data on the joint distribution of two or more variables within each group are usually missing. Therefore, an ecological study may be considered to be based on an incomplete design.

Socioeconomic problems are now seen as health problems that must be addressed to ensure that everyone has an equal chance for a healthy life. By following the Townsend philosophy of relative deprivation and its extension to population level on an ecological scale, a European Deprivation Index (EDI) was proposed by two French team in 2012.

Association between the socio-economic status and the cancer incidence is modelled using Bayesian approach rather than a classical Poisson regression because we expect to encounter overdispersion defined as variability in the number of cases to be higher than expected by the Poisson distribution. The differences in population sizes of the geographical units, called unstructured spatial heterogeneity, might introduce variations and this method permits the distinction between random fluctuations and true variations in the incidence rates. Moreover, neighbouring areas may not be independent and have similar incidence rates. This is called spatial autocorrelation and is also integrated in the Bayesian approach. Therefore, the overdispersed Poisson model was expanded by including spatially dependent and spatially independent random variables and treated with Bayesian approach. We used the hierarchical convolutional Bayes model adding set of explanatory variables for an individual geographical unit that is empirically obtained and corresponding regression coefficients.

2.9.5. Clustering

Geographical regions in which a certain cancer occurs more often can be distributed quite randomly within the geographical area in question, but they can be assigned to groups. Clear groups of geographical units with increased risk can be observed visually. In geographical analyses, we distinguish between methods that search for individual groups (clusters) and clustering methods, i.e. methods of spatial autocorrelation, which is a global index for the entire study area and does not tell us how many groups are in the area and where these are (Waller, Gotway 2004).

For non-smoothed maps Moran's I is provided in the Word output for assessing clustering of the entire geographical area. Moran's I statistics takes values between -1 and 1. Values around zero indicate a random geographical distribution of the observed variable. Moranov I has negative values (towards -1) when the sample is scattered but ordered. When neighbouring geographical units have similar values, Moran's I has a value close to 1, so we can interpret that there are one or more clusters. We assume that the values of the Moran's I statistics are distributed normally. The statistical characteristic was thus checked by Z-test and p-value is also reported.

Moran's I is reported for following figures:

- Observed incidence (Figure 2),
- Age-standardized incidence rate (ASR) (Figures 3a, 3b and 3c),
- The standardized incidence ratio (SIR) (Figure 5).

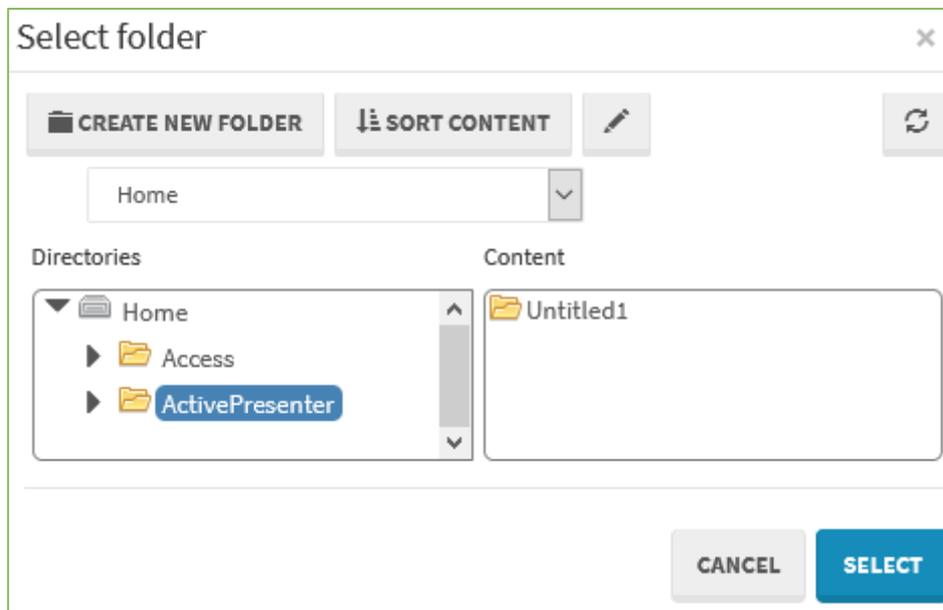
Where Bayesian hierarchical modelling was implemented the clustering was assessed using the ratio between the two precision τ_s / τ_h , with which we can estimate which of the random components has a greater influence on the estimation of the posterior probability distribution. If $\tau_s / \tau_h < 1$, then the spatial random variable S is more important (since its variability is smaller), otherwise the heterogeneous random variable H is more important.

Ratio τ_s / τ_h is reported for following figures:

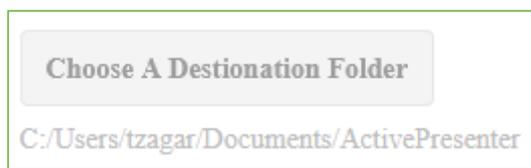
- The deprivation index EDI (Figure 8),
- Smoothed SIR without EDI: INLA (Figure 9).
- Smoothed SIR with EDI: INLA (Figure 10),
- Smoothed SIR without EDI: Gibbs (Figure 11),
- Smoothed SIR with EDI: Gibbs (Figure 12).

2.10. OUTPUT PREFERENCES

First, a destination folder for output files must be selected by clicking the button 'Choose a destination folder'.



After saving the selected folder for outputs is written under the button:



Two output files are always produced:

- Output files are saved in the destination folder after modelling is run by clicking the 'Create output' at the bottom.
 - Output files are not opened automatically.
 - Files are named by the formula (parameters in italics are taken from section 'Basic information'): *Title-wasaby-export-Date*
 - After Word file is saved it is not tackled by the CanMapTool any more
 - After rerunning the models CanMapTool overwrites the outputs in case the files have the same names (when the Title and the date are the same). In case the user wishes to save previous results one has several options:
 - change the Title (in section 'Basic information')
 - rename saved outputs
 - move saved outputs to some other folder
 - choose other destination folder (in section 'Output preferences')
 - The user can change the output files as one wishes.

- In case one of the files is opened while CanMapTool intends to rewritten it, it produces an error (this is only in case user does not specify different name for saving the files – ‘Title’ in the section ‘Basic information’).
- Word file with maps and some calculations (for example Moran I) is a primary output and is always produced.
- Txt file called ‘error_list.txt’ is produced only in case of errors:
 - This file is rewritten every time the output is generated with errors (but if no errors are reported, the file is not generated). User can not see the history of errors unless one saves it intentionally to some other place.
 - This file gives the list of errors but this functionality is still under development and is not much useful to the user. The reason it that the errors are the ones produced by code written in R Statistical Software and are usually not in clear language. The errors are more useful for the developer than for the user of the CanMapTool. Furthermore, the errors do not necessary show the origin of the problem, for example, if something is wrong with uploaded shapefile, the reported error is that plotting is not possible.
- Excel file with numeric outputs is produced only in case the option ‘export .xlsx file with data’. This is useful in case:
 - The user can investigate the exact numbers and not only the colours on the map (where only categories are mapped).
 - The results could be used in some other programme to produce maps or some other analysis.



After the models are run for the first time, two additional options appear:

- Option ‘I want to repeat modelling’:
 - When it is selected all the modelling is calculated again, also those that is already calculated in previous sessions.
 - In case this option is not selected, the performance is much faster. Only modelling for additional models is performed that was not included in the previous session.
- Option ‘uploaded data is the same as in a previous session (population data, incidence data and EDI)’:
 - In case this option is selected, the performance is much faster.
 - This option should be selected in case the user wishes that only new additional models are calculated or if there was any other change in settings (like changing the colour pallets).

- Performance of these two settings is also described with button called 'Clear all data and set settings to default values' under the section 'Creating output'.

I want to repeat modelling

export .xlsx file with data

uploaded data is the same as in a previous session (population data, incidence data and EDI)

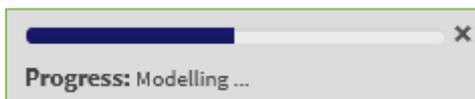
2.11. CREATING OUTPUT

In the last section there is the button '**Create output**', which starts the calculations and drawing the maps. Duration takes several minutes and depends on

- computer's characteristics (memory, processor speed, graphics card , ...)
- numer of regions and
- the selected maps in the 'Map settings' section.



After the button 'Create output' is clicked the progress is indicated in the right bottom corner:



In case something is wrong, CanMapTool produces an error and notifies the user about it. For example in case:

- uploaded data file are not in correct format,
- ID of the regions are not unique or they do not match in all data files,
- the error shown in the CanMapTool is always the same regardless of what exactly went wrong.



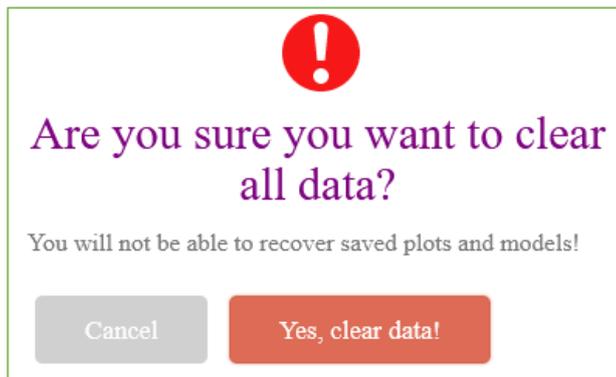
In following situations CanMapTool does not produce an error, but only skips the map, which should be produced in output, or does not produce output at all:

- error at modelling,
- error at drawing maps,
- error at creating output.

There is additional button called **'Clear all data and set settings to default values'**. Performance of CanMapTool and characteristics of this button are:

- This button works in combination with options 'I want to repeat modelling' and 'uploaded data is the same as in a previous session (population data, incidence data and EDI)' under the section 'Output preferences'.
- When user closes the CanMapTool and opens it another time (on the same computer), the CanMapTool preserves all the selected datasets, settings, models and maps. This way the additional outputs are produced faster, because CanMapTool does not calculate everything from beginning.
- This might be useful in scenario as in this example:
 - at the beginning the user only wants to inspect the incidence and population data (checks only option 'Observed incidence by region' and 'Population by region' in section 'Map settings'),
 - after correcting the errors in the datasets the users reloads the incidence and population data (but does not need to reload the shapefiles),
 - only after incidence and population data work appropriately the user can move to ASR and SIR mapping,
 - we recommend adding data on covariates and performing the smoothing only later in the process of analysing the data.
 - Step by step adding the outputs helps the user to carefully investigate the data and results and, in addition, it is recommended for users with slow computers or/and with use on datasets on large areas with high number of regions.
- The button called 'Clear all data and set settings to default values' is to be used:
 - When the user wants to reset all the datasets and settings.
 - It is necessary to use in case of changing any of the following: shapefile, incidence or population data.
 - It is necessary to use in case of analysing different dataset (for example on other area).
 - It is recommended to use in case of replacing one of the datasets (for example on covariate) in case we are not sure if the change was really implemented.

- After clicking the button the user needs to confirm the action by additionally clicking 'Yes, clear data'.



3. DISCLAIMER AND LIMITATIONS OF THE CANMAPTOOL

Cancer registries have responsibility to regular monitor, publish and communicate cancer burden indicators – the basic ones are incidence, prevalence, mortality and survival – in addition to indicators of data quality itself. Not all cancer registries have personnel to additionally regularly perform advanced statistical modelling including geographical analysis, which can not be automated but requires more control by the analyst.

3.1. *DISCLAIMER*

CanMapTool was developed and is available for free use with the intention to promote mapping cancer data, which are available in cancer registries. As such, it tackles the ability of cancer registries to overcome the shortage of knowledge and resources to perform the mapping of the data. But the complexity of mapping itself is the very reason the CanMapTool should not be a full and only substitute for targeted detailed geographical analysis performed by a skilled epidemiologist and should not be taken as such. CanMapTool can be used to perform basic descriptive mapping of the cancer data and for some more complex preliminary data analysis (smoothing and including one covariate). For example, when applying the smoothing techniques (floating weighted averages method and Bayesian hierarchical modelling) the performed models are fixed and can not be modified by the user (except in the source code).

In general, at this stage of the development of the CanMapTool the possibilities for modifications of the modelling parameters by the user are very limited. This is also the reason, we can not guarantee the CanMapTool will perform adequately in every possible data set. Furthermore, it is the user's responsibility to prepare the required datasets with correct content and in the exact format as demanded and described in the CanMapTool instructions. At this stage of the development, it was not possible to implement also comprehensive report on possible errors produced by R code (in which CanMapTool is programmed) so it is up to the user to find the reason why CanMapTool is not working on his dataset (with the help of the list of errors, which is also produced by CanMapTool).

The correct use and interpretation of the results provided by CanMapTool are responsibility of the final user and not the authors of the CanMapTool.

3.2. *GEOCODING*

Mapping in general can not be a substitute for eliminating the problems arising in preparing the datasets or in geocoding spatial attributes. However, it can be used for verifying the data with aim to identify and correct the possible errors. The researcher must inspect the original input data (also, for example, incidence map for small areas) before continuing to produce maps of more complex

outputs such as age-standardised indicators, smoothing, etc., otherwise one can miss errors and weirdness (that are hidden in the end result).

3.3. *SMALL AREAS*

CanMapTool is a general tool and the user can use it at any desired level of areal division. When CanMapTool is used on territorially small-sized regions the colouring might not be visible. For this purpose the

- ‘Map settings’ section offers mapping with hidden region lines.

The other possibilities for zooming in (usually dense populated areas cities with large population but small visible area) are, if the user

- changes the inputs into the CanMapTool to map only desired regions (provides shapefile only for selected regions and corresponding datasets) or
- uses the numerical output from CanMapTool in Excel (in ‘Output preferences’ section there is an ‘export .xlsx file with data’ option for this purpose) and map the required selection by themselves.

CanMapTool can be used with (very) small regions but, still such use is on the user’s responsibility. Using small regions rises several issues, which should be considered before communicating the results. A non-exhaustive list of those includes:

- Smoothing can produce wrong (or at least hard to interpret) results near the border of the whole area.
- Smoothing should not be used across the borders. The borders could be state or country borders (covered by different cancer registries) or borders between subareas with different characteristics influencing the cancer incidence (for example different implementation of the cancer screening programme).
- All smoothing methods implemented in CanMapTool are based on provided data for regions and should be interpreted accordingly.
- Analysis with included covariate EDI (or any variable containing quintiles) are based on provided data for regions and should be interpreted accordingly (also in terms of ecological fallacy).

3.4. *SMALL NUMBERS*

Small-area observations are based on small number of cases and therefore the maps from which anyone can easily estimate the number of cases (in certain age group and gender) should be strictly used for in house investigation only – the applicable data protection legislation has to be considered strictly before publishing such maps. Following the ethics and law is responsibility of the user and not of the authors of the CanMapTool.

4. CLOSING REMARKS

Cancer maps are important tools in public health research. Mapping can be viewed as a descriptive presentation of the cancer burden in some geographical area. They can help to point out areas where health policy should be improved or/and where more detailed analytical research is needed. They are also used for evaluating the performance of public health interventions, like organized screening programs. In any case, maps must be designed to communicate effectively among public, health researchers and decision makers. The biggest challenge is to ensure that maps can not be misinterpreted.

Geographical analyses are feasible when outcomes or exposures or a combination of both have a spatial structure. The use of spatially referenced data in cancer studies is gaining in prominence, fuelled by the development and availability of spatial analytic tools and the broadening recognition of the linkages different data sources. Studies of this nature can assist in public health decision-making. In particular, geographical analyses of the distribution of risk factors can be useful in prioritizing preventive measures. Disease mapping is useful for health service provision and targeting interventions if avoidable risk factors are known.

Geographical studies of disease and environmental exposures may in some cases be sufficient by themselves to justify action, for example if the exposure-disease association is specific, the latency is short and the exposure is spatially defined. Geographic analyses with no information at the individual level are vulnerable to bias. However, while individually based epidemiological studies are in general needed to demonstrate the causal nature of an exposure-disease association, geographical analyses can help strengthen the available evidence.

In the WASABY project, several European cancer registries contributed their datasets for purpose of geographical analyses and mapping. Using the same procedures for all datasets gives great opportunity to compare and point out possible issues one may expect when starting with geographical analyses themselves. CanMapTool targets cancer registry's personnel trying to start up the geographical investigation of their registry's data by themselves.

CanMapTool was developed and is available for free use with intention to promote mapping cancer data, which are available in cancer registries. As such it tackles the ability of cancer registries to overcome the shortage of knowledge and resources to perform the mapping of the data. But the complexity of mapping itself is the very reason the CanMapTool can not be a full substitute for detailed geographical analysis and should not be taken as such. CanMapTool can be used to perform basic descriptive mapping of the cancer data and for some more complex preliminary data analysis (smoothing and including one covariate), the interpretation of the results is still in the hands of cancer epidemiology experts.

5. LITERATURE

1. Australian Cancer Atlas (<https://atlas.cancer.org.au>). Cancer Council Queensland, Queensland University of Technology, Cooperative Research Centre for Spatial Information. Version 09-2018. Accessed 26th of August 2019.
2. Bell BS, Hoskins RE, Pickle LW, Wartenberg D (2006). Current practices in spatial analysis of cancer data: mapping health statistics to inform policymakers and the public. *Int J Health Geogr* 5:49.
3. Bernardinelli L, Clayton D, Pascutto C, Montomoli C, Ghislandi M et al (1995). Bayesian analysis of space-time variation in disease risk. *Stat Med* 14:2433-2443.
4. Besag J (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *J Roy Stat Soc Ser B* 36:192-236.
5. Besag J, York J, Mollie A (1991). Bayesian image restoration with two applications in spatial statistics. *Ann Inst Statist Math* 43:1-59.
6. Breslow NE, Day NE (1987). *Statistical Methods in Cancer Research. Vol. II, The Design and Analysis of Cohort Studies* (IARC Scientific Publication No. 82). Lyon, France: International Agency for Research on Cancer.
7. Brooks S, Gelman A (1998). General Methods for Monitoring Convergence of Iterative Simulations, *J Comput Graph Statist* 7:434-455.
8. Bryere J, Dejardin O, Launay L, Colonna M, Grosclaude P, Launoy G (2018). French Network of Cancer Registries (FRANCIM) Socioeconomic status and site-specific cancer incidence, a Bayesian approach in a French Cancer Registries Network study. *European Journal of Cancer Preventio* 27(4):391-398.
9. Colonna M, Sauleau EA (2013). How to interpret and choose a Bayesian spatial model and a Poisson regression model in the context of describing small area cancer risks variations. *Revue d'E' pide' miologie et de Sante' Publique*, 61:559-567.
10. Dean CB, Ugarte MD, Militino AF (2001). Detecting interaction between random region and fixed age effects in disease mapping. *Biometrics*, 57:197-202.
11. dos santos Silva I (1999). *Cancer Epidemiology: Principles and Methods*. World Health Organization; 2Rev Ed edition.
12. Glattre E, Finne TE, Olesen O, Langmark F (1985). *Atlas of cancer incidence in Norway 1970-79*. The Norwegian Cancer Society, Oslo.
13. Guillaume E, Pornet C, Dejardin O, Launay L, Lillini R, Vercelli M et al (2016). Development of a cross-cultural deprivation index in five European countries. *J Epidemiol Community Health* 70:493-9.
14. Kelsall JE, Diggle PJ (1998). Spatial variation in risk of disease: a nonparametric binary regression approach. *Appl Statist* 47:559-573.
15. Leroux BG, Lei X, Breslow N (2000). Estimation of Disease Rates in Small Areas: A new Mixed Model for Spatial Dependence. In: Halloran ME, Berry D (eds) *Statistical Models in Epidemiology, the Environment, and Clinical Trials*. The IMA Volumes in Mathematics and its Applications, vol 116. Springer, New York.

16. Lokar K, Žagar T, Zadnik V (2019). Estimation of the Ecological Fallacy in the Geographical Analysis of the Association of Socio-Economic Deprivation and Cancer Incidence. *Int. J. Environ. Res. Public Health* 16(3):296.
17. Marmot M, Allen J, Bell R, Bloomer E, Goldblatt P, Consortium for the European Review of Social Determinants of H, et al (2012). WHO European review of social determinants of health and the health divide. *Lancet* 380:1011-1029.
18. Martino S, Riebler A (2019). Integrated Nested Laplace Approximations (INLA). (Submitted on 2 Jul 2019)
19. National Cancer Registry/Northern Ireland Cancer Registry (2011). All-Ireland Cancer Atlas 1995-2007. Cork/Belfast.
20. Pascutto C, Wakefield JC, Best NG, Richardson S, Bernardinelli L, Staines A, et al (2000). Statistical issues in the analysis of disease mapping data. *Stat Med* 19(17–18):2493-2519.
21. Patama T, Pukkala E (2016). Small-area based smoothing method for cancer risk mapping. *Spatial and Spatio-temporal Epidemiology* 19:1-9.
22. Pornet C, Delpierre C, Dejardin O, Grosclaude P, Launay L, Guittet L, et al (2012). Construction of an adaptable European transnational ecological deprivation index: the French version. *J Epidemiol Community Health* 66:982-989.
23. Pritzkeleit R, Eisemann N, Richter A, Holzmann M, Gerdemann U, Maier W, Katalinic A (2016). Krebsatlas Schleswig-Holstein. Räumliche Verteilung von Inzidenz, Mortalität und Überleben in den Jahren 2001 bis 2010. Institut für Krebs Epidemiologie e.V.
24. Pukkala E, Söderman B, Okeanov A, Storm H, Rahu M, et al (2001). Cancer atlas of Northern Europe. Cancer Society of Finland, Helsinki.
25. R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at <http://www.R-project.org/>.
26. Rezaeian M, Dunn G, St Leger S, Appleby L (2007). Geographical epidemiology, spatial analysis and geographical information systems: a multidisciplinary glossary. *J Epidemiol Community Health* 61:98-102.
27. Richardson S, Thomson A, Best N, Elliott P (2004). Interpreting Posterior Relative Risk Estimates in Disease-Mapping Studies, *Environ Health Perspect* 112:1016-1025.
28. Riebler A, Sørbye SH, Simpson D (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25(4):1145-1165.
29. Rue H, Martino S, Chopin N (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Statist. Soc. B* 71(2):369-392.
30. Simpson DP, Rue H, Martins TG, Riebler A, Sørbye SH (2017). Penalising model component complexity: A principled, practical approach to constructing priors (with discussion). *Statistical Science*, 32(1):1-46.
31. Townsend P (1987). Deprivation. *J Soc Pol*, 16:125-146.

32. Waller LA, Gotway CA (2004). Applied Spatial Statistics for Public Health Data. John Wiley & Sons, Inc, New Jersey.
33. Žagar T, Zadnik V, Primic Žakelj M (2011). Local standardized incidence ratio estimates and comparison with other mapping methods for small geographical areas using Slovenian breast cancer data. Journal of Applied Statistics, 38(12):2751-2761.