WASABY

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

D4.3
Cancer incidence and water & soil pollutants: Data management model to perform spatial analysis

WP4: Roberto Lillini

Working group: Martina Bertoldi, Fortunato Bianconi, Maurizio Zarcone, Tina Žagar, Elodie Guillaume, Joséphine Bryere, Ludivine Launay, Alessandro Borgini, Paolo Contiero, Paolo Baili

V1: 31st December 2019

WASABY European Project – WP4 Report

D4.3 – Cancer incidence and water & soil pollutants: Data management model to perform spatial analysis

Table of contents

Co-funded by
the Health Programme
of the European Union

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

1. Report aims

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 the 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 additional health resources, such as cancer control programs, interventions and program dissemination.

WASABY aims to design a model able to identify areas with higher cancer rates, so to study whether pollutant contamination may be a cause for increased cancer risk.

The main focus of this Report is to provide an operative model designed and validated for developing incidence rates maps. This report uses information coming from a literature review on studies performing spatial analysis, as well as from the activities conducted in WP4, WP5 and WP6. It is prepared utilizing breast cancer data collection as an example of how CRs can deal with geo-coding, use of socio-economic data, spatial analysis. This report will deal with:

- the characteristics of data geo-coding, giving information on how this task could be performed affordably, according to the information available in a Cancer Registry (CR) database;
- the choice and role in spatial analysis of covariates which are not collected from health institution, such as the socio-economic deprivation indices;
- the characteristics and reasons behind the choice of one or more statistical model for spatial analysis;
- an hint to the implementation of the environmental pollution covariates in such kind of analysis, considering the next steps of the WASABY project as a reliable example.

In synthesis, this report explains the steps to follow by a Cancer Registry to perform environmental spatial analysis of incidence data for water and soil pollutants, taking advantage of the previous WASABY reports prepared in the first phase of the project.

2. LITERATURE REVIEW ON PAST ENVIRONMENTAL SPATIAL ANALYSIS STUDIES

During 2019, we performed a PubMed literature review aimed to individuate past studies useful as example for WASABY objectives. In fact, our review aimed to identify various environmental study methodologies applied across the world to evaluate correlation between water and soil pollutants (e.g., arsenic in water, topsoil metals, etc.) and a given health outcome (e.g., cancer incidence, acute gastrointestinal infection hospital admissions, etc.) with data available independently from the aim of the environmental study as in WASABY.

The PubMed search identified 694 articles. After three phases of article revision, 40 articles were included in the review and classified.

A scientific article was recently submitted with the results of this literature review.

Four types of methods were identified by the review:

1. Regression models using data at individual level
2. Regression models using data by geographical areas
3. Computation of threshold values for exposure intensity, in order to define cut-off points for evaluating trends in the health outcome variable influenced by the environmental factor
4. Comparisons between health outcomes geographic clusters and environmental pollution geographic clusters by considering the distance between them

While most of the considered works shared the cross-sectional study design as expected, the analysis of the articles in this review has shown the wide variation of valid and reliable methods and techniques used for investigating these phenomena. It was not possible to identify a “gold standard”, because of the peculiarities of every situation and of the data available for environmental pollutants and/or health outcomes.

However, starting from Cancer Registry incidence data, the best models to be applied are the “Regression models using data at individual level or by geographical areas”.

For this reason, in the present report we list the various steps to follow for performing these types of analysis with CR data.

The following paragraphs indicated the suggested steps to be followed by a CR to perform environmental spatial analysis with own incidence data:

- Data management:
  - Design the study protocol and Authorization Request to the Ethical Committee
  - Data collection:
    a. incidence and population data by same age group, geographical unit and calendar year;
    b. geocoding and shapefiles for geographical units that are valid for the corresponding time period;
    c. socio-economic data (SES data) and other possible cofounders;
    d. environmental data.
  - Preparation of the final dataset:
    a. linkage of cancer and background population data on the same geographical scale;
    b. classifying dataset according to availability of geocoded data into case A, B or C (see paragraph 4.1) and allocating residence addresses to the desired geographical units for each cancer case if appropriate;
    c. data quality check;
    d. linkage with covariates not directly collected by the CR, as the socio-economic information (e.g., the European Deprivation Index - EDI),
    e. production of a final geo-coded dataset.
- Spatial analysis and smoothing methods
- Spatial analysis adjusting for covariates (SES, environmental data, etc)

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

3. DATA MANAGEMENT

3.1 DESIGN THE STUDY PROTOCOL & REQUEST TO THE ETHICAL COMMITTEE

Designing the protocol, a CR must take in consideration that three groups of data are needed for preparing cancer incidence maps:

1) cancer cases;
2) background population;
3) geospatial vector of geographical territory.

Data on cancer cases and background population should be georeferenced to the same geographical level. Geographical information on residence has to be added to the cancer patients' data by the registry. For this reason, the available level of geographical units can vary considerably. Considering this, the participating CRs are classified according to availability of geocoded data:

- Case A: cancer cases with information on residence address but not geocoded at x and y level.
  - Investigation is needed to determine whether it is possible to allocate the desired geographical units of the residence address to each cancer case.
  - The dataset becomes case B if allocation is possible.
- Case B: Cases with geocoded residence address but not linked to desired sub-area.
  - Geocoded residence addresses will be allocated to the desired geographical units for each cancer case.
  - Analyses and mapping for aggregated data will be performed.
- Case C: Cases with geocoded and linked residence address to desired sub-area.
  - Point-based addresses will be allocated to desired geographical units.
  - Analyses and mapping for aggregated data will be performed as in case B.
  - Analyses for point data will be performed but only if background population is also available at point level (or coordinates of reasonable approximation).

In the preparation of the protocol, we have also to consider that geographical analyses and mapping differ in respect of 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 classed choropleth maps are the standard.

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

- Point data: exact coordinates of residence address (or coordinates of reasonable approximation) for cancer incidence cases and 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 also geographically changing age structure of population should be taken into account (for example using local SIR estimates or calculating risk surface by generalized additive model).

Together with cancer cases and background population CRs also provided shapefiles for geographical units that are valid for corresponding time period. The shapefile format is a popular geospatial vector data format for geographic information system (GIS) softwares and is usually provided by national mapping authorities.

Official shapefiles (i.e., coming from official sources as the ISTAT – Italian National Statistics Office) should be used, in order to allow reliable and reproducible geocoding, mapping and analysis. Usually, official national and local statistics office can provide the required files; shapefiles are composed by at least three different files to five files:

- A .shp file, which contains the graphical information for mapping.
- A .dbf file, which contains the database with all the geographic information converted into data.
- A .shx file, which contains a shape or font compiled by Autodesk AutoCAD from an .SHP shape file or .PFB font file. It stores shape definitions, as well as font definitions for displaying custom text.
- A .prj file, which reports the characteristics of the used geographic projections (e.g., WGS84).
- A .qrj file, that is another file reporting projection information.

Obviously CRs are responsible for data completeness in both terms (i.e., complete case ascertainment and information for each case). Missing information for specific cases (such as missing age or address) must be rechecked by the CR. These are the issues of CR's data quality.

CRs are also responsible for (re)coding the addresses into geographical units for the analysis of aggregated data. In case a CR provides exact addresses with x and y coordinates (point data) for cancer cases, but the population data is on the aggregated level, the cases can be allocated (using GIS function and shapefiles) to the provided geographical units (same as population data is given) for purpose of ASR and SIR calculation on aggregated level.

Data management software like Excel, STATA and R can be used for data management. Geographic software like ARCGis or QGis can be used for the check of the shapefiles.

Co-funded by the Health Programme of the European Union

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

3.2 DATA COLLECTION: INCIDENCE AND POPULATION DATA[1]

Incidence is an absolute number of all newly diagnosed cases of any disease in a defined population in one calendar year. Data on population covered by the CR are also necessary. 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 diagnosed with more than one different cancers.

The incidence registered by CRs 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,
- address.

Other data describing cancer entity, personal characteristics or environment may be additionally included as covariates.

Possible issues with provided datasets are:

- In case a CR provides exact addresses with x and y coordinates (point data) for cancer cases, but the population data is on the aggregated level, the cases will be allocated (using GIS function and shapefiles) to the provided geographical units (same as population data is given) for purpose of ASR and SIR calculation on aggregated level.
- When incidence and population datasets cover different calendar years, only overlapping years contained in both datasets will be included into the analysis.
- In case there is some geographical unit with no female population, the unit will be joined with neighbouring unit having the smallest female population for purpose of map presentation only. Geographical unit with no female population will be excluded from analysis.
- In case dataset includes a calendar period of more than 10 years, the size of geographical units will be checked in terms of number of cases and population in smallest units. If reasonable, the relevant CRs will be asked to produce several maps for shorter time periods (but not less than a 5 years span). This because ten years is the most common period at the end of which changes in geographic areas can be found and new Census relevation are performed. Such changes can, obviously, affect the incidence computation significantly.

Software like Excel, Stata and R can be used for this kind of data management, due to their flexibility and customizability.

[1] The text and content of 3.2, with some emendations, come from the WP6 Report: "D6.1 - WASABY Report on methods v2", authored by the people at WP6.

Co-funded by
the Health Programme
of the European Union

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

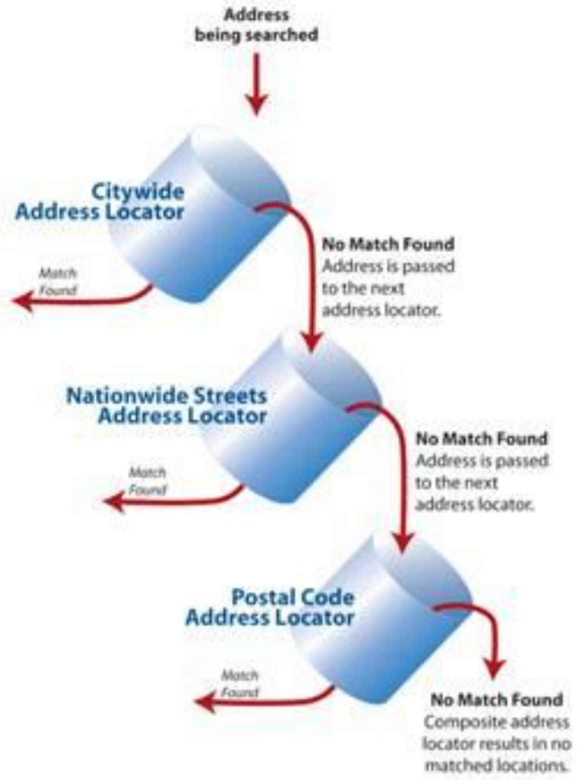
3.3 DATA COLLECTION: GEOCODING[2]

CRs 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 CRs, mainly because of different situations specific for each country, which influence accessibility of detailed or coded address. Addresses can be available in different forms (for example coded or alphanumerically transcript) and levels – both impact on the possible analyses selection.

Geocoding is the process of converting addresses (e.g., a street address) into geographic coordinates (e.g., 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 y = 49.203529, x= -0.354513 in WGS84 coordinate system, corresponding to French census tract IRIS number 141181404. Geocoded addresses can be subsequently aggregated to any geographical units (administrative or user defined), which is beneficial when dataset cover longer time periods that are subject to changes in geographical units.

Figure 1: Illustration of the geocoding process.



Source: http://help.arcgis.com/fr/arcgisdesktop/10.0/help/index.html#//00250000003r000000

[2] The text and content of 3.3, with some emendations, come from the WP6 Report: "D6.1 - WASABY Report on methods v2", authored by people at WP6.

Co-funded by
the Health Programme
of the European Union

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

The preparation for geocoding needs the following steps to be completed:

- The required information is a precise address – registered by CRs – 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 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 or collectively (https://qgis.org/en/site/forusers/download.html). Other programmes include qVSIG or GRASS GIS (the list is not exhaustive). As QGis is widely used, there are many tutorials and discussions on the internet.
- To make the link between address and geographical maps are needed. Some maps are commercialized by famously established editors (for example, ESRI) but the price can be expensive according to the product one needs. Free data are available on openstreetmap.

The geocoding process is performed step by step (Figure 1). Address is used with all information (number, type, name, postal code, and 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).

The following steps compose the geocoding process:

- preparation of address formatting street name, locality (to optimize automatic geocoding);
- quality control of original data;
- preparation of data allowing to differentiate the process according to locality type (more or less than 5,000 inhabitants) in GIS (often programming in Python language);
- geocoding (mainly correction of addresses that have no correspondence);
- allocating the corresponding chosen administrative unit for each coordinates (e,g., IRIS in France, Census Tract in Italy, etc.);
- geocoding quality control: addresses automatically matched are in accordance with original locality;
- adding geo-coded covariates (e.g., SES data) according to geographical unit.

The already-cited geographic software (ARCGis, QGis) can perform most or all of the geo-coding procedures. The Excel free-service add-on Excel Geocoding Tool can be used to recover correct address and x and y coordinates for each case.

3.4 Data collection: SES data

Both individual and environmental factors contribute to the risk of cancer and the prognosis for affected patients. In cancer epidemiology, the impact of SES on incidence rate and prognosis is increasingly recognized (Woods et al., 2006). Concerning prognosis, the direction of influence of socioeconomic deprivation is always the same, regardless geographical area and cancer site, the most unfavourable prognosis being always the prerogative of most deprived populations. Concerning incidence, the direction depends on cancer site, certain cancers, as head and neck, being more frequent in deprived people, others, as melanoma, being more frequent in affluent people (Bryere et al., 2017). Such impact may be seen through direct effect, for example through increased exposure to environmental hazards in low-income areas, or through indirect effect by creating stressors that enhance the impact of chemical, physical, and lifestyle and behavioural factors that influence cancer incidence. Therefore, one of the potential social confounders to consider in cancer incidence spatial analysis is a contextual deprivation index to characterize socioeconomic neighbourhood level. For studies including different European countries, we suggest to use the European Deprivation Index (EDI) as indicated in WASABY WP-5 reports.

As a general consideration, the CRs should check in their country/area the availability of the socio-economic information (both as single variables or aggregated in a deprivation index) at the same geographic unit of the geo-coded cases and population, in order to perform the matching between the two different kind of information.

For instance, in Italy the socio-economic information coming from the National Census, the National Deprivation Index and the EDI are available at census tract level. Therefore, if a CR would like to use such variables as covariates/confounders for evaluating to the risk of cancer and the prognosis for affected patients, their residence addresses should be geo-coded at census tract level, in order to perform the correct linkage.

3.5 Data collection: Other confounders

Not only SES influences the incidence risk of cancer and the prognosis. Individual factors, e.g. ethnicity, family history, age, reproductive factors, alcohol intake, weight, physical activity, hormone therapy and oral contraceptives, have been found to influence, for example, the risk of breast cancer. Adherence to organized screening programmes in areas covered by cancer registries, lead to an increment of incidence in those areas. The major issue is that these data, however, are not always available at individual level, but they could be very interesting variables to study the differences in incidence and prognosis.

Where possible, all these information could be collected at census-block level, but the CRs should accurately check their availability and reliability. Availability and reliability issues in these kinds of variables could be related to: non-systematic collection; non-certified processes of collection and check; inhomogeneities of the sources (e.g, public vs. private, etc.); incomplete anagraphic information (fundamental for linkage); privacy issues.

3.6 DATA COLLECTION: ENVIRONMENTAL DATA

Official environmental data sources are often available online. Local, national and international open source databases can be considered to individuate correct data source for environmental spatial analysis. In this report we refer only to water&soil pollutants. A list of European databases is reported in Table 1 as individuate by WASABY WP7.

These databases can be divided in two categories:

- databases collecting water and soil quality measures;
- databases collecting pollution emission values due to the presence of point sources, such as industrial sites, landfills and other productive activities.

Table 1 – Main European Environmental Databases

Name	Argument	Web address	Organization	Included Countries *	Time period
Waterbase - Water Quality	Water quality data	https://www.eea.europa.eu/data-and-maps/data/waterbase-water-quality-2	EEA's databases	38	2000 - 2016
FOREGS Geochemical Atlas of Europe	Soil quality data	http://weppi.gtk.fi/publ/foregsatlas/	FOREGS	26	1998 - 2002
LUCAS TOPSOIL	Soil quality data	https://esdac.jrc.ec.europa.eu/content/lucas-2009-topsoil-data	JRC	28	2009-2012
E-PRTR	Emission Data of Industrial sites	https://prtr.eea.europa.eu/#/home	EEA's databases	33	2007-2017
IPCHEM Information Platform for Chemical Monitoring	Air, Water, soil, quality data	https://ipchem.jrc.ec.europa.eu/RDSIdiscovery/ipchem/index.html	Different EEA's databases and other databases	38	1980- 2017

Source: WP7 Report: "D7.2 - Report on environmental data available for spatial analysis".

Furthermore, the European platform (IPCheM) provides a systematic collection of environmental databases produced at local, national and European level, in the context of European projects or in compliance with national or European environmental regulations.

Co-funded by
the Health Programme
of the European Union

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

Both the databases collecting measured concentration data or measured emission data also contain exposure maps to contaminants obtained through the use of statistical-mathematical models described each time in each database.

The main issues of the environmental data is the geographic level at which they are available: in most of cases they are provided by sporadic point sources of emission, collected only in specific moments in time, and not necessarily for purposes of connecting them with epidemiological study (more frequently, they are collected for monitoring the quality of the environment).

Therefore, before using the environmental information in a study like WASABY, a CR must cope with some relevant issues:

- The proven relationship between the considered pathology/ies and the collected pollutants.
- The quality of the information (which subject collected such data?) and the time consistence of the data with the considered outcome.
- The area covered by the collected environmental data and what statistical model can be the most appropriate to estimate the pollutant distribution across the area corresponding to the geographic unit chosen for the study (e.g., census tract).

The WASABY review, cited in paragraph 2 of this report, give more detailed information on his topics.

3.7 Estimation measures: Age standardized incidence rates (ASR)[3]

Crude cancer incidence rate (crude rate) is defined as number of new cases (O) in a specified time period, divided by the number of persons, living in observed area specific population (pop_i) in the same time interval and geographical unit (denoted with index i):

$$CrudeRate_i = O_i / pop_i$$

Crude rate is usually expressed as the number of cancers per 100,000 population at risk. Since the datasets consist of cases with age at diagnosis of cancer up to 49, crude incidence rate is not an appropriate measure for investigating and presenting geographical distribution. Age standardization is used in epidemiological analyses since it takes into account 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

[3] The text and content of 3.8, with some emendations, come from the WP6 Report: “D6.1 - WASABY Report on methods v2”, authored by people at WP6.

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.

Age-specific rates (R_a) in 5-year age groups (index a), starting at 15 years (i.e., 15-19, 20-24,..., 45-49) are calculated for investigated time period for each geographical unit. These age-specific rates equal the number of cases divided by the corresponding population. Age-standardised rates (ASR_i for geographical unit i) are calculated by multiplying the age-specific rates (R_{ai}) by standard population weights (N_a) and then adding together.

$$ASR_i = \sum_a R_{ai} \cdot N_a \,.$$

Age-specific standard deviation is given as

$$SD_{ai}^2 = O_{ai} / n_{ai}^2 \,.$$

In calculating standard deviation of ASR_i for geographical unit i, also population's distribution is accounted in:

$$\mathrm{SD}_i^2 = \sqrt{\sum_a (\mathrm{SD}_{ai} \cdot \frac{\mathrm{N}_{ai}}{\mathrm{N}_i})^2}$$

And confidence interval (CI) is given by (α is confidence level, $z_{\alpha/2}$ is the (100·α/2) the centile of the standard normal distribution):

$$\mathrm{CI}_i = \mathrm{ASR}_i \pm \mathrm{z}_{\alpha/2} \cdot \mathrm{SD}_i$$

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.

European or World standard population can be used for age standardization, for purpose of comparisons between countries.

Usually, ASR values for geographical units are categorised into classes for mapping purpose using standard deviation. Map of ASR are produced for given geographical units using a colour palette for purpose of stressing the differences.

R, Stata, ArcGIS and QGIS can be used for analyses and mapping.

3.8 Estimation measures: Standardized incidence ratios (SIR)[4]

We assume the observed number of new cancers (O_i) in each single geographical unit i follows a Poisson distribution with mean $\mu_i = E_i\theta_i$, where θ_i is the unit specific relative risk. E_i is the expected number of new cases if the population in a particular area (pop_i) has the same age-specific incidence rates as some larger comparison population (R'_a for age group a in reference population), usually the overall population of the whole study area, or some other reference population. E_i is derived from indirect standardization. Observed and expected numbers of cases can be compared, because both refer to same population. The ratio of the observed number of cases to that expected is called standardized incidence ratio (SIR):

$$SIR_i = O_i / E_i \,;$$

$$E_i = \sum_a R'_{ai} \cdot pop_{ai} \,.$$

Confidence interval for each geographical unit i is given by Fisher's exact test as (χ is from chi-squared distribution, α is confidence level):

$$\mathrm{CI}_{i,lower} = \frac{\chi^2_{\alpha/2,2\mathrm{O}_i}}{2\ \mathrm{E}_i}$$

$$\mathrm{CI}_{i,upper} = \frac{\chi^2_{(1-\frac{\alpha}{2}),2(\mathrm{O}_i+1)}}{2\ \mathrm{E}_i}$$

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. This means SIR maps cannot 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. A ratio less than 1 indicates a lower than average relative risk and over 1 is a higher than average. The variance of θ_i is proportional to E_i-1 and so, for areas with small population size, there will be high sample variability in geographical units with small population.

As with the direct method, the result depends in part upon the standard chosen. 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. For reference population, the overall population of the whole study area is used and age-specific incidence rates are calculated.

Usually, SIR values for geographical units are categorised into classes for mapping purpose. Map of SIR are be produced for given geographical units using a colour palette for purpose of stressing the differences.

R, Stata, ArcGIS and QGIS can be used for analyses and mapping.

[44] The text and content of 3.9, with some emendations, come from the WP6 Report: "D6.1 - WASABY Report on methods v2", authored by people at WP6.

3.9 DATA MANAGEMENT: THE FINAL DATASET

Finally, the CR should be able to connect and present all the data coming from the different sources; more specifically, to obtain such dataset a CR should have performed:

a. Linkage of cancer and background population data on the same geographic scale. This information should report the cancer cases at the same geographic unit, along with other information (e.g., gender, age, date of incidence, tumour characteristics), chosen for the study. Quite often, the geographic scale is the smallest geographic and administrative unit represented in a shapefile (e.g., the census tract). Therefore, in the data matrix the units of analysis are the geographic units which aggregates the patients' and population information. If such information is available at individual level, then the units of analysis would be the geo-coded patients.

b. Linkage with the geographic variables in the shapefiles, including the geometry, the cartography and the geographic characteristics of the area. The linkage must be performed at the same level of the previous step and it's necessary for the spatial analysis and representation.

c. Data quality check, in terms of homogeneity of the linked information.

d. Computation of ASR and/or SIR at the geographic scale at which the dataset has been built.

e. Linkage with covariates not directly collected the CR, as the socio-economic information (e.g., the European Deprivation Index - EDI).

f. Linkage with the environmental variables after adjustment for the same geogragriphic scale of the cancer cases.

The final dataset should be a data matrix where:

- The columns/variables report all the data regarding patients, population, geograpich elements, other not-CR-collected information, and pollutants concentration data.
- The rows are the units of analysis and represent the geographic scale at which the analysis will be performed, more or less disaggregate according to the shared data availability.

Co-funded by
the Health Programme
of the European Union

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

4. SPATIAL ANALYSIS AND SMOOTHING

Geographical units are problematic in terms of their size and the population they cover. 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 four 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.

4.1 SARAR MODELS FOR SPATIAL ANALYSIS

A first considered approach is an inferential one: the Spatial Autoregressive with Autoregressive Disturbance model (SARAR).

The SARAR model is based on the Cliff - Ord model which, in its simplest version, considers only the spatial effects in the dependent variable, with the effects modelled including a covariate known as a spatial "lag". Each observation of the spatial "lag" variable is a weighted average of the values of the dependent variable observed for the other units of the cross section.

The matrix containing the weights is known as the spatial weighting matrix. This model is often referred to as the autoregressive spatial model (SAR). A generalized version of this model also allows evaluating the disturbances generated by a SAR process.

The SAR model combined with SAR disturbances is often referred to as a SARAR model.

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

In modelling the outcome for each geographical unit of analysis as dependent on a weighted average of the results of other units, the SARAR models determine the results simultaneously. This diversity implies that the ordinary least squares estimator will not be consistent.

SARAR models are implemented in Stata and R software. Here we present some solution driven by the Stata version of the algorithm.

In Stata, the *spreg* command implements a maximum likelihood estimator (ML) and a generalized two-stage least squares spatial estimator (GS2SLS) for the parameters of a SARAR model with exogenous regressors. The following notation will be used: for each matrix A and vector a, the elements are indicated as a_{ij} and a_i respectively.

The *spreg* command estimates the parameters of the cross-sectional model in an extended and compact version:

$$\begin{aligned} y_i &= \lambda \sum_{j=1}^{n} w_{ij} y_j + \sum_{p=1}^{k} x_{ip} \beta_p + u_i \\ u_i &= \rho \sum_{j=1}^{n} m_{ij} u_j + \varepsilon_i \end{aligned}$$

$$\begin{aligned} \mathbf{y} &= \lambda \mathbf{W} \mathbf{y} + \mathbf{X} \beta + \mathbf{u} \\ \mathbf{u} &= \rho \mathbf{M} \mathbf{u} + \epsilon \end{aligned}$$

Where:

- y is the n × 1 vector of the observations of the dependent variable;
- W and M are the n × n weighted spatial matrices (with 0 diagonal elements);
- Wy and Mu are n x 1 vectors defined as spatial lags;
- λ and ρ are the corresponding scalar parameters, defined as SAR parameters;
- X is the n x k matrix of the observations relating to the k exogenous variables (covariates), where some variables can represent the spatial lag of the exogenous variables);
- β is the corresponding k × 1 vector parameter;
- ε an n × 1 vector of innovations (stochastic effects).

The two previous models are two forms of a SARAR model with exogenous regressors, that is, the definition of which is not linked to the dependent variable but exclusively to the choice of the geographical units of analysis.

Co-funded by the Health Programme of the European Union

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

Spatial interactions are modelled through spatial lag. The model allows spatial interactions in the dependent variable, in exogenous variables and in disturbing effects.

Spatial weighting matrices W and M are considered known and not stochastic. These matrices are part of the model definition and, in many applications, W = M.

$$\overline{y}_i = \sum_{j=1}^{n} w_{ij} y_j$$

The previous formulation shows the dependence of y_i on the results of the neighbouring areas, through the spatial lag yi. By construction, the Wy spatial delay is an endogenous variable. The w_{ij} weights will typically be modelled as inversely related to a certain measure of proximity between the units. The SAR λ parameter measures the extent of these interactions.

The stochastic effects ε are independent and identically distributed (IID) or independent but heteroschedastically distributed, where heteroskedasticity is of unknown form.

The GS2SLS estimator produces consistent estimates in both cases when the heteroskedastic option is specified. The ML estimator produces consistent estimates in the IID case but generally not in the heteroskedastic case.

The model presented above is a first order SAR model with first order SAR disturbances and is also referred to as the SARAR model (1, 1). It is a special case of the more general SARAR model (p, q).

When ρ = 0, the model boils down to the SAR model:

$$y = \lambda W y + X\beta + \varepsilon.$$

When λ = 0, the model is reduced to: y = Xβ + u, with u = ρMu + ε, sometimes referred to as the SAR error model.

Setting ρ = 0 and λ = 0 causes the model to be reduced to a linear regression model with exogenous variables.

The *spreg* command requires that the spatial weighting matrices M and W be provided in the form of an spmat object; *spreg gs2sls* supports both general spatial weighting matrices and limited spatial weighting matrices; *spreg ml* only supports general matrices.

Spatial weighting matrices are used to model the interactions between spatial units or, more generally, between cross-sectional units. The spatial lag of a variable is defined as the weighted average of the observations on the variable with respect to neighboring units. The n × 1 vector Wy is generally referred to as the spatial lag in y and the n × n matrix W as the spatial weighting matrix.

$$\mathbf{W} = \begin{bmatrix} 0 & w_{12} & \cdots & w_{1,n-1} & w_{1n} \\ w_{21} & 0 & \cdots & w_{2,n-1} & w_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ w_{n-1,1} & w_{n-1,2} & \cdots & 0 & w_{n-1,n} \\ w_{n1} & w_{n2} & \cdots & w_{n,n-1} & 0 \end{bmatrix}$$

More generally, the concept of spatial lag can be applied to any variable, including exogenous variables and disturbances. Spatial weighting matrices allow us to conveniently implement Tobler's first geographical law: "everything is related to everything else, but neighboring objects are more related than distant ones", which applies if the space is geographic, biological or social.

Finally, in the SARAR models the smoothing is based on the moving averages method, applied to generalized least squares averages.

4.2 Floating weighted averages method[5]

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 (as in the approach of this WP6). Floating weighted averages aim at diminishing the random variation by locally calculating floating averages, weighted by population (population weights are denoted by w_i^{pop}):

$$w_i^{pop} = \frac{pop_i}{pop},$$

[5] The text and content of 4.2, with some emendations, come from the WP6 Report: “D6.1 - WASABY Report on methods v2”, authored by people at WP6.

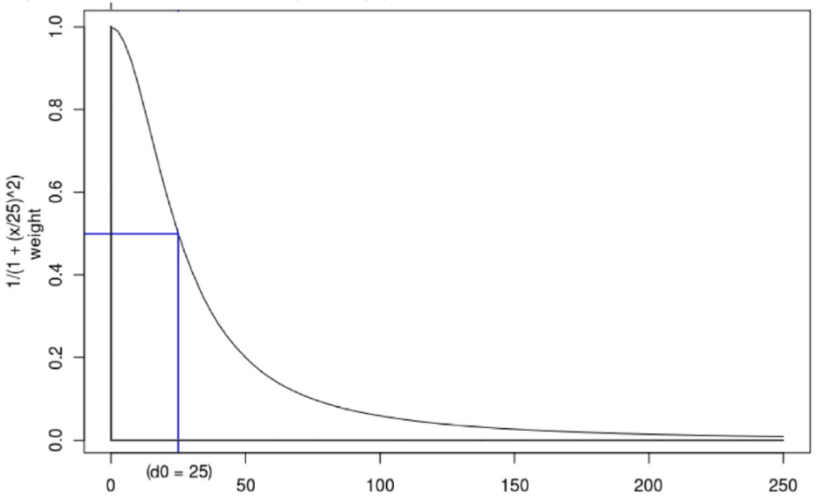
Co-funded by the Health Programme of the European Union

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

where the whole population (pop) is the sum of area specific population (pop_i). Simultaneously we weight by distance (d) using the following formulae:

$$w_i^d = \begin{cases} 1/\left(1+\left(d/D_0\right)^{2m}\right) ; & d \leq D_{max} \\ 0 ; & d > D_{max} \end{cases} .$$

Figure 2: Distance weighting function.



In most cases, the maps seem to smooth nicely when radius D_{max} is about 10-fold compared to distance weight factor D_0. Although the maximum distance of influence (D_{max}) seems quite large, one should bear in mind that the spatial weights are diminishing fast by distance. Further extension of D_{max} normally results in visible changes on the map, whereas finding an optimal value for d_0 may require iterations. If d_0 is short, the single geographical unit's specific rates (in this case, mainly small geographical units with zero cases) will become visible. With setting D_0 =100 km, the rural areas have lost almost all variation and we are close to a situation of over-smoothing.

Finally, both spatial and population weights are multiplied:

$$w_i = w_i^d \cdot w_i^{pop} .$$

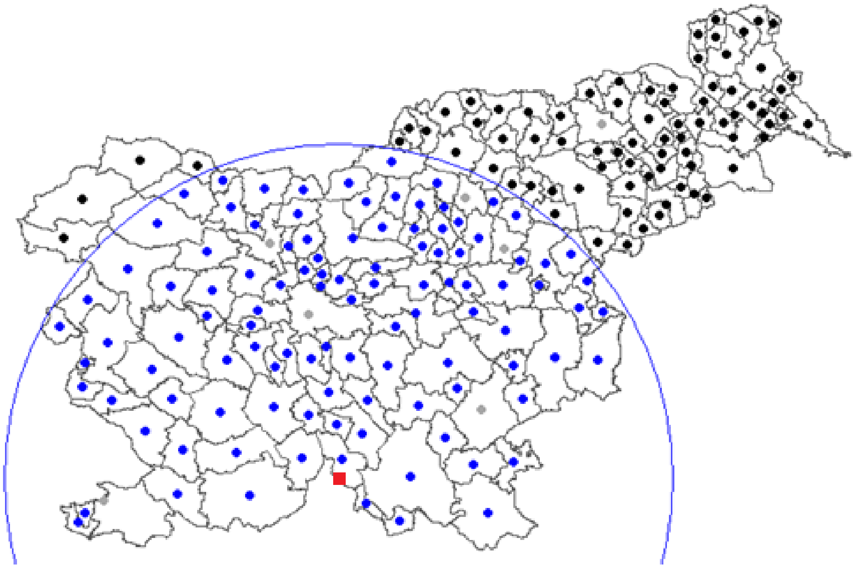
Because of population weighting, large cities (that is, municipalities which include large cities) significantly influence its neighbourhood when smoothing. It has been shown that cancer burden can vary between the main cities and the surrounding less urbanised regions. For this reason, selected big 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.

For observed geographical unit the grid net needs to be constructed and the calculated SIR for each grid point (indexed by g) is then:

$$SIR_g = \frac{\sum_{i'} SIR_{i'} w_{i'}}{\sum_{i'} w_{i'}} .$$

In the last equation index i' runs over the selected geographical units except through excluded larger cities.

Figure 3: Description of floating weighted averages method on Slovenian municipalities' case.



Blue dots are the centroids for selected municipalities used for calculation of smoothed standardized incidence ratio (SIR) for one selected grid point g (SIRg). The blue circle has the maximum radius (Dmax). Red square marks the selected grid point g. Grey dots are big cities excluded from calculation of smoothed picture and are positioned at the centroids of the principal cities themselves rather than at the centroids of the corresponding municipalities. Black dots are neither big cities nor in the Dmax range of the selected grid point.

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

In synthesis:

- The grid areas will be constructed with 500m distance covering the whole geographical areas.
- Starting parameters for preparing maps with floating weighted averages method are:
    - Geographical units with more than 20,000 female population will be excluded from smoothing and their rates shown in circles above the smoothed background.
    - D_0 = 20km, D_{max} = 200km and m = 1.
    - After investigation of the resulting maps, the parameters will be adjusted and new maps prepared. The same parameters will be used when preparing ASR and SIR maps with Finnish smoothing method for the same dataset.
- Two maps will be prepared:
    - ASR (and accordingly SIR) values for grid points will be categorised into classes for mapping purpose. Grid points are coloured instead of assigning the color to the whole geographical unit's area.
    - ASRs from paragraph 2.2 smoothed with floating weighted averages method. For all dataset, the same classes and colour palette will be used as in paragraph 2.2.
    - SIRs from paragraph 2.3 smoothed with floating weighted averages method. For all datasets, the same classes and colour palette will be used defined in paragraph 2.4.
- R and ArcGIS will be used for analyses and mapping.

4.3 BAYESIAN HIERARCHICAL MODELLING[6]

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.

WinBUGS and BYM

The convolution model originally proposed by Besag et al.:

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

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

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. The CAR model with L2 norm (also called a Gaussian Markov random field) for S has an improper density

$$p(S \mid \tau_s) \propto \tau_s^{(n-G)/2} exp(-\frac{\tau_s}{2} S'QS) ,$$

where τ_s controls smoothing induced by this prior, larger values smoothing more than smaller ones; G is the number of “islands” (disconnected groups of regions) in the spatial structure; and Q is n × n with nondiagonal entries q_{ij} = - 1 if regions i and j are neighbours and 0 otherwise, and diagonal entries q_{ii} are equal to the number of region i’s neighbours. This is a multivariate normal kernel, specified by its precision matrix $\tau_s Q$ instead of the usual covariance matrix. 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.

[6] The text and content of 4.3, with some emendations, come from the WP6 Report: “D6.1 - WASABY Report on methods v2”, authored by people at WP6.

R-INLA and BYM2

WinBUGS and MCMC have long been used for Bayesian hierarchical modelling. In the classic model of Besag, York and Mollié, BYM, 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. 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. The posterior distribution will be approximated by using the Gibbs sampler in WinBUGS software. Running two independent Markov chains are recommended. The 'burn-in' samples need to be discarded (for example, we will start by discarding first 10,000 out of total 20,000 iterations). Convergence of relative risk will be confirmed by graphing their traces and observing random mixing of chains, which revealed white noise variation around a common value with no trend. This was supported by observing Brooks–Gelman–Rubin diagnostics, which clearly satisfied convergence criteria.

WinBUGS and BYM have long been used for Bayesian hierarchical modelling as a gold standard, so we will use it for comparison in the scope of WASABY project. However, since this statistical field has recently developed, we will also apply modified methodologies for Bayeasian hierarchical modelling: R INLA and BMY2. For Bayesian hierarchical modelling R INLA package and WinBUGS software will be used. R and ArcGIS will be used for other analyses and mapping.

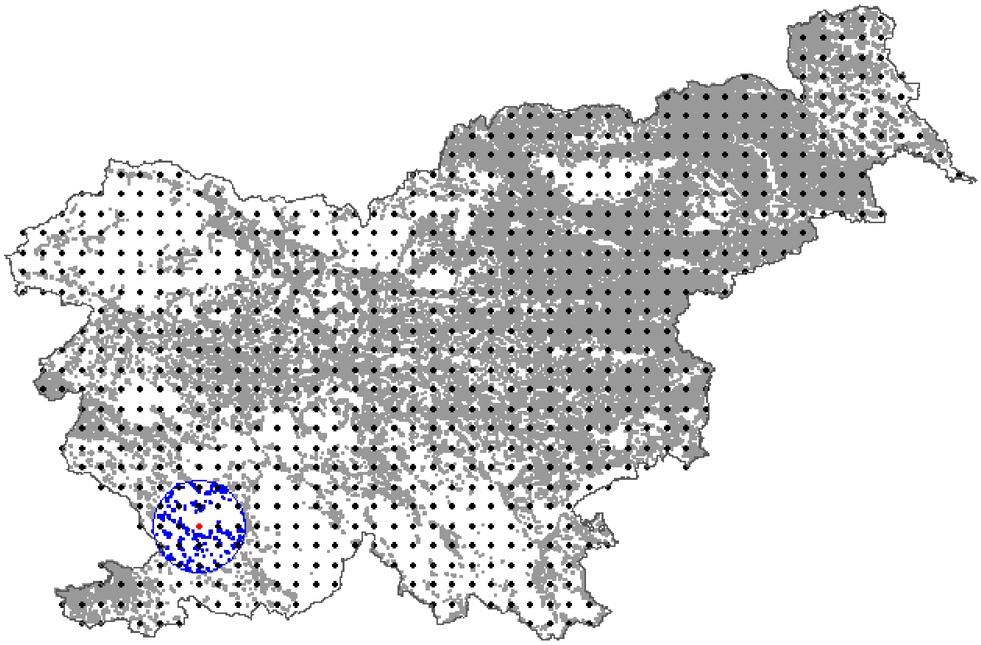
EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

4.4 LOCAL SIR ESTIMATES[7]

Local SIR estimates use circular “moving” window centred on grid locations covering the whole study area. SIRs is calculated for each grid point. The procedure requires data on exact x, y coordinates of the residence of each person included into the analysis with indication, which are also the cancer cases. For each person, the information about gender and age group is needed for the purpose of (indirect) age-standardization. At each fine grid location, the circle is centred and case and population data occurring within the circle are determined. Based on this information, SIR is calculated, which belongs to this specific grid point.

The circle radius is not fixed in advance but is changing from D_{min} to D_{max} with predefined step until predetermined minimum population is reached. This way, for each grid point, the calculated SIRs are based on (approximately) the same number of persons at risk giving more stable estimates. D_{min} can be chosen equal or larger than the grid spacing in order for circles to overlap. D_{max} ca should not be larger than half of the width or height of the study area. D_{max} controls maximum distance of influence.

Figure 4: Description of local standardized incidence ratio (SIR) estimates method on Slovenian case.



[7] The text and content of 4.4, with some emendations, come from the WP6 Report: “D6.1 - WASABY Report on methods v2”, authored by people at WP6.

Co-funded by the Health Programme of the European Union

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

Black dots: grid points. Red dots: the selected grid point for which the SIR is calculated using only the data within circular blue window (which is moving through grid points). Blue colour: the population that fall into the selected window. Grey colour: the rest of the population not included in the window.

Generally, the approaches based on exact geographical locations have problems with estimates near area border. The grid points at the border have missing population and increasing the circle does not account for that and usually generates biased estimates. One solution would be to reduce the circle radius or the population criteria in the border. However, this operation increases the variance of the estimates. Further on, at the border there can be also very sparse population. Therefore, it is not recommended to map the values for grid points, where the minimum population requirement is not reached in circle with radius D_{max} (the map in such parts is not coloured).

Local standardized incidence ratio estimates allow to use point data for preparing cancer maps in fine resolution, thus revealing more localized patterns and ignoring the arbitrary administrative borders. The map of the local SIR estimates emphasizes extremes, but unlike the map, based on the observed SIRs, these estimates are stable, enabling more accurate evaluation. The disadvantage is that the geocoded data are not routinely available.

Summarising these procedures:

- The map of local SIR estimates will be prepared for datasets whe both population and cancer incidence datasets will have geocoded data to an x and y coordinates of the residence.
- The same set of grid points will be used as in the floating weighted averages method.
- Starting parameters for preparing maps with floating weighted averages method are:
  - D_{min} = 1km, D_{max} = 15km with step 1km until predetermined minimum population is reached.
  - Minimum population requirement is 5,000.
  - After investigation of the resulting maps, the parameters will be adjusted and new maps prepared. Same parameters will be used when preparing ASR and SIR maps with Finnish smoothing method for the same dataset.
- SIR values for grid points will be categorised into classes for mapping purpose.
- R and ArcGIS will be used for analyses and mapping.

Co-funded by the Health Programme of the European Union

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

5. Spatial analysis: adjusting for covariates

Ecological analysis is defined as the assessment of the associations between disease incidence (e.g., suicide) and variables of interest (e.g., 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, the EDI was proposed as a measure of socio-economic inequalities comparable between different countries.

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:

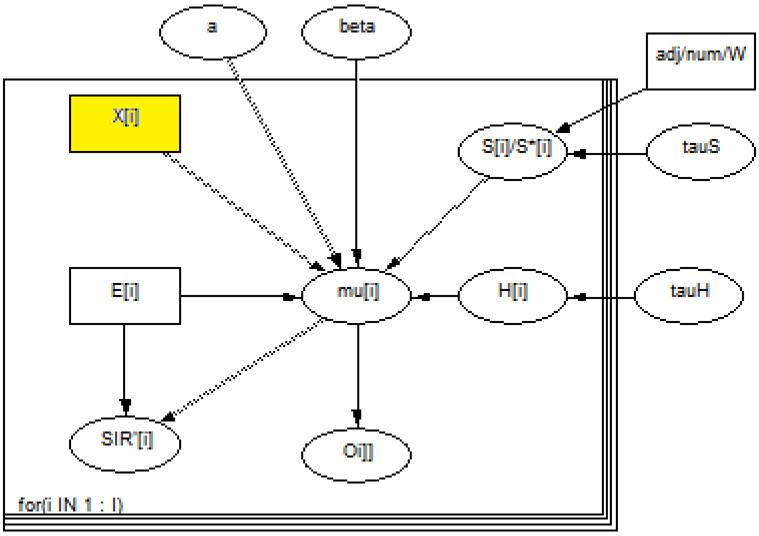
$$\ln\frac{O_i}{E_i} = \ln E_i + a + \sum_{j=1}^{J} \beta_j x_{ij} + H_i + S_i$$

where a represents the basic (logarithmic) relative risk of disease in the entire study area. O_i and E_i represent the observed and expected number of cases in the i-th geographical unit. H_i are unstructured (heterogeneous) random factors that are geographically independent and S_i is a spatially dependent component (spatially structured heterogeneity). We define it by a conditionally autoregressive (CAR) prior probability distribution. X_{ij} is a set of J explanatory variables for an individual geographical unit that is empirically obtained. β_j are the regression coefficients for the j-th explanatory variable. The model is the same as in paragraph 3.2 with added fixed effect X_{ij}, in our case EDI.

Co-funded by
the Health Programme
of the European Union

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

Figure 5: An example of a model with explanatory variables (X_i) constructed in WinBUGS.



The regression coefficient associated with the variable EDI and its 95% confidence interval are estimated in the model. A positive parameter related to EDI means an over-incidence in disadvantage areas and a negative parameter related to EDI means an over-incidence in affluent areas.

This step will be conducted only for datasets, where European deprivation index (EDI) is provided by CR for the same geographical units as incidence and population data. For countries where EDI is not available, local available deprivation indices geo-coded at the same geographic level (e.g., census tract) will be used.

EDI (or other socio-economic deprivation index) dataset will be linked to the cancer incidence dataset, using census tract as the linking variable. EDI will be classified into quintiles, meaning each dataset (country) will have different classes for EDI. R, ArcGIS and WinBugs will be used for analyses and mapping.

Co-funded by
the Health Programme
of the European Union

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

6. Conclusions

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 cannot be misinterpreted.

Geographical analyses are feasible when outcomes or exposures or a combination of both have a spatial structure. 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 CRs 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.

7. Bibliography

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 Statis 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. Drukker DM, Prucha IR, Raciborski R. (2013) (b) Maximum likelihood and generalized spatial two-stage least-squares estimators for a spatial-autoregressive model with spatial-autoregressive disturbances. The Stata Journal; 13(2): 221-242.
13. Drukker DM, Peng H, Prucha IR. (2013) (b) Creating and managing spatial-weighting matrices with the spmat command. The Stata Journal; 13(2): 242-286.

Co-funded by
the Health Programme
of the European Union

EU funding disclaimer: This project has received funding from the 3rd European Union Health Programme 201-2020 under Grant Agreement PP-2-5-2016 (# 769767)

14. Glattre E, Finne TE, Olesen O, Langmark F (1985). Atlas of cancer incidence in Norway 1970-79. The Norwegian Cancer Society, Oslo.
15. Guillaume E, Pornet C, Dejardin O, Launay L, Lillini R, Vercelli M, Marí-Dell'Olmo M, Fernández Fontelo A, Borrell C, Ribeiro AI, Pina MF, Mayer A, Delpierre C, Rachet B, Launoy G. Development of a cross-cultural deprivation index in five European countries. J Epidemiol Community Health. 2016 May;70(5):493-499
16. Kelsall JE, Diggle PJ (1998). Spatial variation in risk of disease: a nonparametric binary regression approach. Appl Statist 47:559-573.
17. 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.
18. Lillini R, Quaglia A, Vercelli M; Registro mortalità Regione Liguria. [Building of a local deprivation index to measure the health status in the Liguria Region]. [Article in Italian]. Epidemiol Prev. 2012 May-Aug;36(3-4):180-187.
19. Lillini R, Vercelli M. The local Socio-Economic Health Deprivation Index: methods and results. J Prev Med Hyg. 2019 Feb 28;59(4 Suppl 2):E3-E10.
20. 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.
21. Martino S, Riebler A (2019). Integrated Nested Laplace Approximations (INLA). (Submitted on 2 Jul 2019)
22. National Cancer Registry/Northern Ireland Cancer Registry (2011). All-Ireland Cancer Atlas 1995-2007. Cork/Belfast.
23. 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.
24. Patama T, Pukkala E (2016). Small-area based smoothing method for cancer risk mapping. Spatial and Spatio-temporal Epidemiology 19:1-9.
25. 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.
26. Pritzkuleit 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 Krebsepidemiologie e.V.

27. Pukkala E, Söderman B, Okeanov A, Storm H, Rahu M, et al (2001). Cancer atlas of Northern Europe. Cancer Society of Finland, Helsinki.
28. 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/.
29. 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.
30. Ribeiro AI, Launay L, Guillaume E, Launoy G, Barros H. The Portuguese version of the European Deprivation Index: Development and association with all-cause mortality. PLoS One. 2018 Dec 5;13(12):e0208320.
31. Richardson S, Thomson A, Best N, Elliott P (2004). Interpreting Posterior Relative Risk Estimates in Disease-Mapping Studies, Environ Health Perspect 112:1016-1025.
32. Riebler A, Sorbye 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.
33. 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.
34. 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.
35. Waller LA, Gotway CA (2004). Applied Spatial Statistics for Public Health Data. John Wiley & Sons, Inc, New Jersey.
36. Zadnik V, Guillaume E, Lokar K, Žagar T, Primic Žakelj M, Launoy G, Launay L. Slovenian Version of The European Deprivation Index at Municipal Level. Zdr Varst. 2018 Apr 6;57(2):47-54.
37. 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.