



# WASABY

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

## D6.1

### Report on methods for spatial analysis of the cancer data

**WP6:** Tina Žagar

**Working group:** Martina Bertoldi, Fortunato Bianconi,  
Joséphine Bryere, Marc Colonna, Andreja Kuhec, Ludivine Launay,  
Ron Pritzkeleit, Eero Pukkala, Marc Saez, Vesna Zadnik, Maurizio Zarcone

**V2:** 9<sup>th</sup> September 2019

# WASABY European Project – WP6 Report

## D6.1 – Report on methods for spatial analysis of the cancer data

### Table of contents

<b>1. INTRODUCTION AND GENERAL OBJECTIVES OF WP6 .....</b>	<b>3</b>
<b>2. MAPPING CANCER INCIDENCE .....</b>	<b>4</b>
2.1. REQUIRED DATASETS .....	4
2.2. GEOCODING .....	7
2.3. INCIDENCE.....	9
2.4. AGE STANDARDIZED INCIDENCE RATES (ASR) .....	10
2.5. AGE STANDARDIZED INCIDENCE RATIOS (SIR) .....	13
2.6. SCALE AND COLOUR PALETTE.....	14
<b>3. SMOOTHING .....</b>	<b>16</b>
3.1. FLOATING WEIGHTED AVERAGES METHOD.....	16
3.2. BAYESIAN HIERARCHICAL MODELLING.....	19
3.3. LOCAL SIR ESTIMATES .....	22
<b>4. ADJUSTING FOR COVARIATES .....</b>	<b>25</b>
<b>5. CONCLUSIONS.....</b>	<b>27</b>
<b>6. LITERATURE .....</b>	<b>29</b>

## 1. INTRODUCTION AND GENERAL OBJECTIVES OF WP6

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. The present delivery D6.1 is an interim delivery for WP6 final documents: the manual for CRs on mapping cancer burden (D6.2) and spatial analysis on breast cancer in young women applying data of the participating CRs (D6.3). Deliverable D6.1 describes the procedures and methods which will be used in D6.3. D6.1 contents, however, are not binding for D6.3: the WASABY WP6 group of experts may decide at a later stage to change an approach, a method or a computer programme. D6.2 will be broader, covering also selected additional methods we are not implementing in D6.3.

The main focus of preparing a common report for participating CRs is on visualization and also on promoting reliable maps in both methodological and epidemiological terms. For this reason, smoothing techniques and adjusting for covariates are also discussed. There are numerous spatial smoothing techniques – we have selected three 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.

## 2. MAPPING CANCER INCIDENCE

Over the latest 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 mostly used in few 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 are rarely routinely available in the cancer epidemiological research. However, spatially referenced public health data sets have become more available in recent years.

### 2.1. REQUIRED DATASETS

Technically, 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. In WASABY, participating cancer registries (CR) provided female breast cancer cases for 10 most recent calendar years, with age at diagnosis up to 49, for the smallest available geographical units. Furthermore, all of the CRs provided background population data for the same calendar years and same geographical units. Geographical information on residency was added to the cancer patients' data within each registry. For this reason, the available level of geographical units 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.
  - 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 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).

Geographical analyses and mapping differ on 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.
- 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.

The data collected from participating CRs were verified within the WASABY team and pre-processed to be suitable for further analysis and final mapping.

### *Wasaby's procedure, part 0*

- CRs provided incidence and background population data by same 5-year age group, a geographical unit and calendar year. CRs also provided shape files for geographical units that are valid for the corresponding time period.
- The data collected from participating CRs were preprocessed to be suitable for further analysis in final mapping. The management of dataset included:
  - linkage of cancer and background population data on the same geographical scale,
  - classifying dataset according to availability of geocoded data into case A, B or C and allocating residence addresses to the desired geographical units for each cancer case if appropriate,
  - data quality check,
  - linkage with the socio-economic information (the European Deprivation Index, EDI),
  - production of a final dataset for each CR.
- CRs are responsible for data completeness in both terms i.e., complete case ascertainment and information for each case. Cases not recorded by CR can not be compensated by modelling in WASABY. Typically, the datasets are prepared to as small geographical units as achievable by each CR, so underreporting may result in misleading outputs. Missing information for specific cases (such as missing age or address) will be rechecked with the CR. These are the issues of CR's data quality and, if detected in course of data analysis, the matter of interpretation of the results.
- CRs are 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 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.
- Excel, STATA and R will be used for data management.

## 2.2. GEOCODING

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 (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 our previous example, geographical coordinates correspond to 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.

WASABY project stimulates participating CRs to include geocoding into their routine registration process.

Preparation for geocoding:

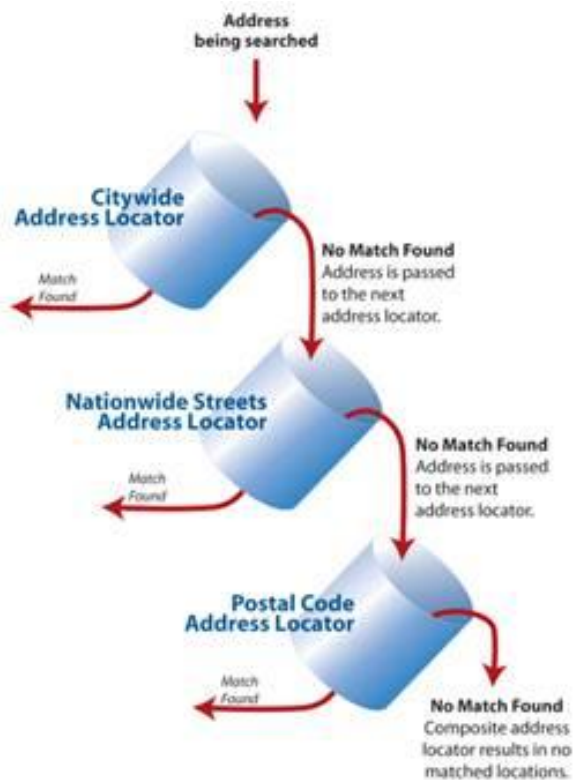
- The required information is a precise address – registered by CRs – including house number, street type, street name, postal code, municipality or city (other code specific to country, for example in France, insee code which is specific for each municipality). Be aware that this information is considered 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 collectivity (<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. This should be the most appropriate for discovering GIS.Maps.

- To make the link between address and geographical coordinates one needs maps.. Some maps are commercialized by famoulsy 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, 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).



**Figure 1:** Illustration of the geocoding process. Source: <http://help.arcgis.com/fr/arcgisdesktop/10.0/help/index.html#//00250000003r000000>.



### *The example of the procedure performed by Caen*

Caen is using ArcGIS 10.7 provided by ESRI France. Maps for geocoding are BD Adresse for ArcGIS Advanced® (provided by ESRI France and IGN).

The following steps compose the geocoding process:

- preparation of address with formatting street name, locality (to optimize automatic geocoding) programming with SAS® 9.4 Software,
- quality control of original data (for example, locality in accordance with insee code) programming with SAS® 9.4 Software,
- preparation of data allowing to differentiate the process according to locality type (more or less than 5,000 inhabitants) in GIS programming in python,
- geocoding (mainly correction of addresses that have no correspondence),
- allocating the corresponding IRIS for each coordinates,
- geocoding quality control: addresses automatically matched are in accordance with original locality as with IRIS,
- adding European Deprivation Index according to geographical unit (mainly IRIS).

## **2.3. INCIDENCE**

Incidence is an 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 he/she is 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 and address. These are also the inclusion criteria for the analyses foreseen in D6.3. Other data describing cancer entity, personal characteristics or environment may be additionally included as covariates.

### Wasaby's procedure, part 1

- 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).
- Excel and R will be used for data management.

## 2.4. AGE STANDARDIZED INCIDENCE RATES (ASR)

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 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.

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

Age-specific standard deviation is given as

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

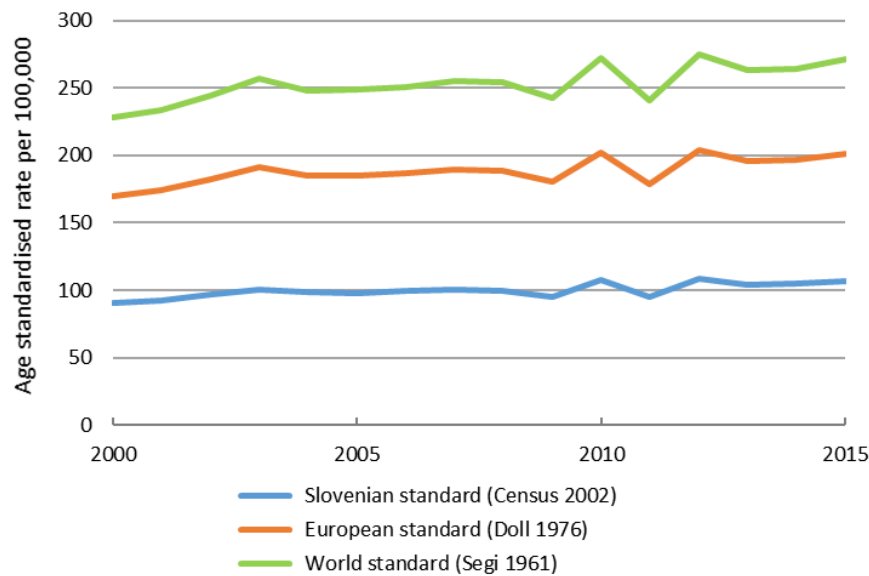
In calculating standard deviation of  $ASR_i$  for geographical unit  $i$ , also population's distribution is accounted in:

$$SD_i^2 = \sqrt{\sum_a (SD_{ai} \cdot \frac{N_{ai}}{N_i})^2}$$

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

$$CI_i = ASR_i \pm z_{\alpha/2} \cdot SD_i$$

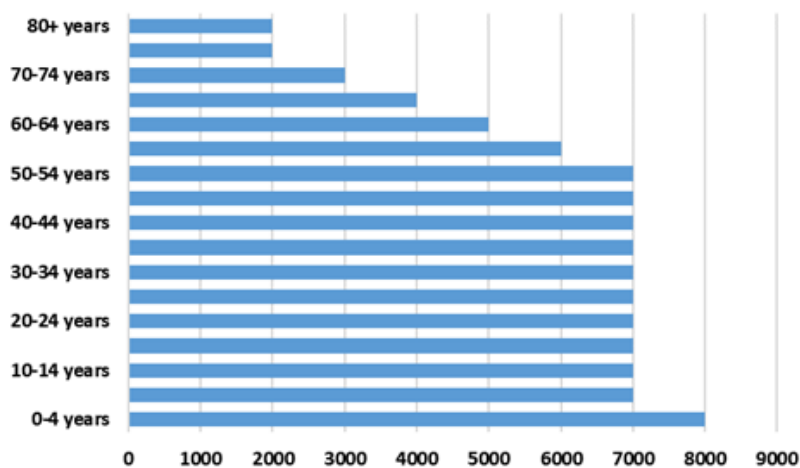
For the most part, the choice of weights (the standard population) is based on convention, the intended and potential comparisons, and various other considerations. There is often no absolute correct choice, and there can easily be different opinions about the best one. Regardless of the chosen standard 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.



**Figure 2:** Comparison of applying different standard populations on breast cancer incidence in females, Slovenia 2000-2016.

### Wasaby's procedure, part 2

- European standard population (Doll 1976) will be used for age standardization for purpose of comparisons between European countries.



**Figure 3:** Population structure in European standard population (Doll 1976).

- ASR values for geographical units will be categorised into classes for mapping purpose using standard deviation.
- Map of ASR will be produced for given geographical units using colour palette in paragraph 2.4.
- R, ArcGIS and QGIS will be used for analyses and mapping.

## 2.5. STANDARDIZED INCIDENCE RATIOS (SIR)

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  $\theta_i$  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 ( $\chi$  is from chi-squared distribution,  $\alpha$  is confidence level):

$$CI_{i,lower} = \frac{\chi^2_{\alpha/2, 2O_i}}{2 E_i}$$

$$CI_{i,upper} = \frac{\chi^2_{(1-\alpha/2), 2(O_i+1)}}{2 E_i}$$

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 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. A ratio less than 1 indicates a lower than average relative risk and over 1 is a higher than average. The variance of  $\theta_i$  is proportional to  $E_i^{-1}$  and so, for areas with small population size, there will be a high sample variability in geographical units with small population.

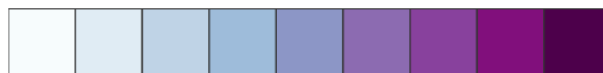
As with the direct method, the result depends in part upon the standard chosen. However, the indirect method of standardization is less sensitive to the choice of standard than the direct one. Indirect method is also preferable to the direct method when age-specific rates in geographical unit is based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

### *Wasaby's procedure, part 3*

- For reference population, the overall population of the whole study area will be used, and age-specific incidence rates will be calculated.
- SIR values for geographical units will be categorised into classes for mapping purpose.
- Map of SIR will be produced for given geographical units using the colour palette in paragraph 2.4.
- R, ArcGIS and QGIS will be used for analyses and mapping.

## **2.6. SCALE AND COLOUR PALETTE**

The aim of the WASABY project is not comparing cancer incidence between countries or regions so there is no need for fixed scaling when mapping ASRs. The sequential colour scheme of nine classes starts with light purple shades for small values starting with 0 and turns dark purple at the highest values, the highest limit being maximum of all ASRs. Areas in which three or fewer cases of disease occurred were given a grey colour due to the in-stability of the data.



**Figure 4:** Sequential colour scheme for mapping ASRs. The colours are chosen from ColourBrewer pallets available at <http://colourbrewer2.org/>.

Relative scaling is used for mapping SIRs. Since SIR is interpreted as the relative risk relative to the national average, a divergent colour scheme is appropriate with quasi-continuous gradient. The middle class is the brightest and presents relative risk equal to national (can be interpreted as national average). Darker and more vivid colours indicate greater deviations from the national average than lighter, paler colours.

All values lower than half and higher than twice the national average fall into the darkest two classes. The next 15 classes (dark green to pale light green) represent the value range 0.5 (= half national average) to 0 (= no deviation from the national average). A normal distribution with mean value 1 and a standard deviation of 0.268 means that approximately the same number of cases may be expected for each class. The area above the national average (brown colour) is reciprocal. The top class in the darkest brown contains all values that are greater than 2, ie twice the national average. The 15 classes between 1 and 2 were formed by the reciprocal value of the correspondent lower classes.

The same scale and colour pallet will be used for all CR's datasets. The same colours in two maps based on the same dataset can be compared (for example SIRs and smoothed SIRs). The smoothed SIRs deviate less from national average, so light colours are used in greater extent when mapping. The same colours in two maps based on different datasets (for example different countries or regions) should not be compared directly, as they correspond to different national average. The relative deviation to the country average for the same colour is, however, equal.



**Figure 5:** Divergent colour scheme for mapping SIRs. The colours are chosen from ColourBrewer pallets available at <http://colourbrewer2.org/>.

### 3. 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 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.

#### 3.1. FLOATING WEIGHTED AVERAGES METHOD

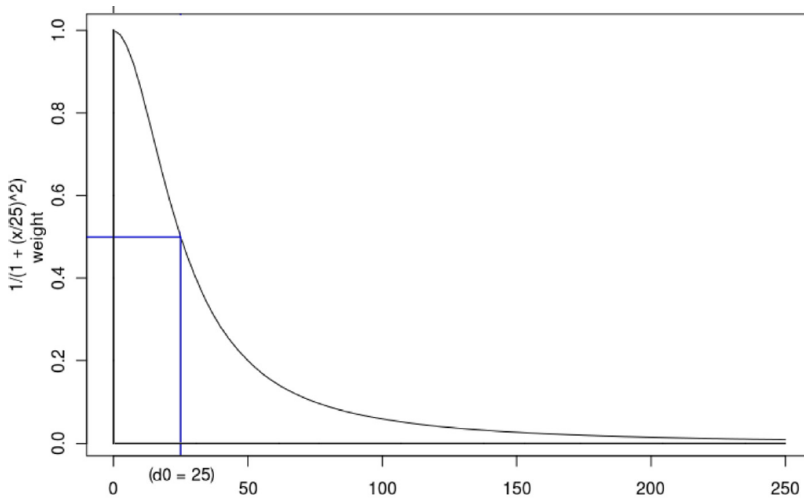
The “Finnish smoothing method” uses floating weighted averages and was first used in the national cancer incidence atlas of Norway and further developed in the Finnish atlas and in the Cancer Atlas of Northern Europe. The floating weighted averages method has mostly been applied to age-adjusted incidence 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  $w_i^{pop}$ ):



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

where the whole population (pop) is the sum of area specific population (pop<sub>i</sub>). Simultaneously we weight by distance (d) using the following formulae:

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



**Figure 6:** 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 7:** 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 (SIR<sub>g</sub>). The blue circle has the maximum radius (D<sub>max</sub>). 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 D<sub>max</sub> range of the selected grid point.

#### *Wasaby's procedure, part 4*

- 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.
  - $D_0 = 20\text{km}$ ,  $D_{\max} = 200\text{km}$  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.

### **3.2. 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.

### 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  $\tau_h$ . 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

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

where  $\tau_s$  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  $n \times n$  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  $\tau_s$  and  $\tau_h$  have Gamma priors (0.5, 0.0005) as suggested by Bernardinelli et al.

### R-INLA and BYM2

WinBUGS and MCMC have long been used for Bayesian hierarchical modelling so we will use it for comparison in the scope of WASABY project. 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.

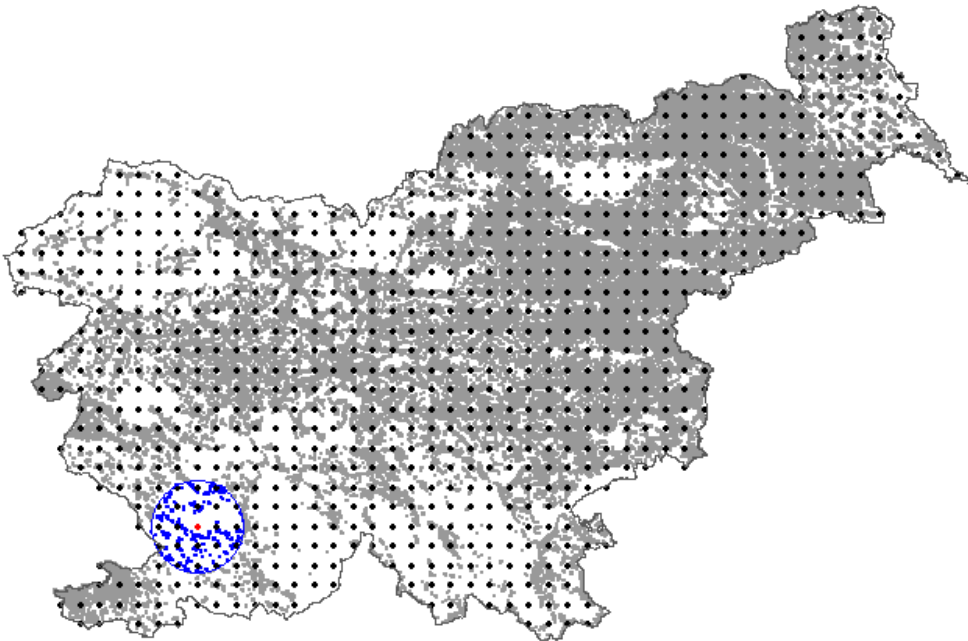
### Wasaby's procedure, part 5

- 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 Bayesian hierarchical modelling: R INLA and BM2.
- For Bayesian hierarchical modelling R INLA package and WinBUGS software will be used. R and ArcGIS will be used for other analyses and mapping.

### 3.3. LOCAL SIR ESTIMATES

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.  $D_{\max}$  should not be larger than half of the width or height of the study area.  $D_{\max}$  controls maximum distance of influence.



**Figure 8:** Description of local standardized incidence ratio (SIR) estimates method on Slovenian case. 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.

### *Wasaby's procedure, part 6*

- The map of local SIR estimates will be prepared for datasets where 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:
  - $D_{\min} = 1\text{km}$ ,  $D_{\max} = 15\text{km}$  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.



#### 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.

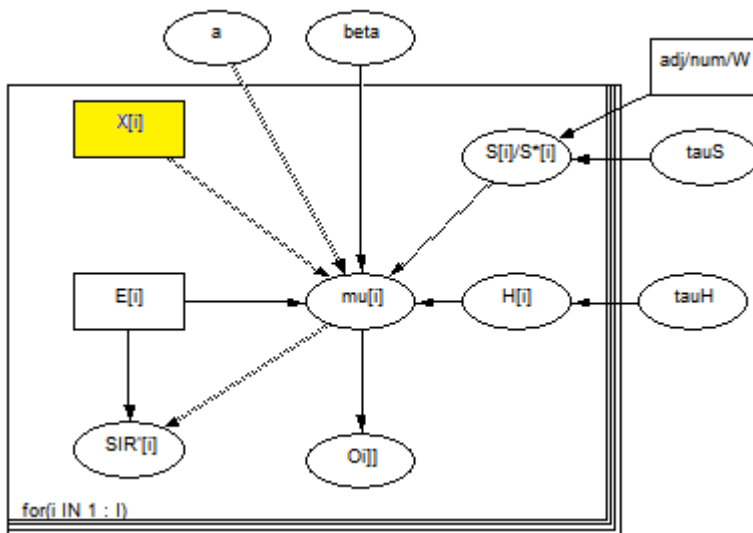
Socioeconomic problems are now seen as health problems that must be addressed to ensure that everyone has an equal chance for a healthy life. By following the Townsend philosophy of relative deprivation and its extension to population level on an ecological scale, a European Deprivation Index (EDI) was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

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

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.  $\beta_j$  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.



**Figure 9:** 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.

#### Wasaby's procedure, part 7

- 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.
- EDI dataset will be linked to the cancer incidence dataset.
- 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.

## 5. 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 can not 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. Joining the experience from all tasks performed in D6.3 will be an important part of the practical manual (D6.2). Such manual is targeted to CR's personnel trying to start up the geographical investigation of their registries' data by themselves. The practical manual will be expanded with examples on participating CRs' datasets and identification of methods and softwares useful for studying geographic distribution of cancer incidence.

### *Wasaby's procedure, part 8*

- Each report for specific dataset will be concluded with:
  - list of performed analyses and prepared maps,
  - some remarks on issues with datasets and
  - specifics with dataset and data manipulation to include in practical manual for CRs (D6.2).

## 6. 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 26<sup>th</sup> 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. Kelsall JE, Diggle PJ (1998). Spatial variation in risk of disease: a nonparametric binary regression approach. *Appl Statist* 47:559-573.

14. 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.
15. 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.
16. Martino S, Riebler A (2019). Integrated Nested Laplace Approximations (INLA). (Submitted on 2 Jul 2019)
17. National Cancer Registry/Northern Ireland Cancer Registry (2011). All-Ireland Cancer Atlas 1995-2007. Cork/Belfast.
18. 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.
19. Patama T, Pukkala E (2016). Small-area based smoothing method for cancer risk mapping. *Spatial and Spatio-temporal Epidemiology* 19:1-9.
20. 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.
21. 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.
22. Pukkala E, Söderman B, Okeanov A, Storm H, Rahu M, et al (2001). Cancer atlas of Northern Europe. Cancer Society of Finland, Helsinki.
23. 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/>.
24. 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.
25. Richardson S, Thomson A, Best N, Elliott P (2004). Interpreting Posterior Relative Risk Estimates in Disease-Mapping Studies, *Environ Health Perspect* 112:1016-1025.
26. 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.

27. 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.
28. 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.
29. Waller LA, Gotway CA (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley & Sons, Inc, New Jersey.
30. 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.