WASABY

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

D7.3 Protocol for Pilot environmental study

Paolo Contiero, Alessandro Borgini, Martina Bertoldi, Andrea Tittarelli, Roberto Lillini, Paolo Baili

V1 – 28th February 2020

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 2014-2020 under Grant Agreement PP-2-5-2016 (# 769767)

PROTOCOL FOR
PILOT ENVIRONMENTAL STUDY FOR THE WASABY PROJECT
V1 – 28th February 2020

1. BACKGROUND

The WASABY project (WAter & Soil contamination and Awareness on Breast cancer risk in Young women) focuses on the geographical analysis of population based cancer incidence data in connection with environmental factors. Project's details are at www.wasabysite.it.
Specifically, WASABY aims to evaluate correlation between water & soil pollutants (e.g. arsenic in water, topsoil metals, etc.) and a given health outcome (i.e. breast cancer incidence) with data available independently from WASABY.

2. WASABY PILOT ENVIRONMENTAL STUDY

The WASABY Pilot Environmental Study is a feasibility study evaluating the relationship between health outcome (i.e. cancer incidence), environmental factors (i.e. water & soil pollutants) and/or any other covariate (i.e. socio-economic indicators) by means of quantitative and qualitative approaches using cancer registry and environmental data already available in national or international databases of pollutants.
The WASABY Pilot Environmental Study will be developed in at least one area covered by the participating cancer registries (CRs) listed in Annex 1.
The present protocol is prepared as *vademecum* for each CR interested in developing a similar study inside its covered area.

3. DATA TO BE COLLECTED

Different data sources could be considered to develop WASABY Pilot Environmental Study:

a) Cancer registries to collect cancer cases to estimate cancer outcome indicators (i.e. cancer incidence). These data can be at individual level (XY coordinates) or at smallest administrative unit (SU). See Annex 2 for details on data to be collected for WASABY aims
b) National offices or directly cancer registries to have data on population number of the areas covered by the cancer registry. These data are at smallest administrative unit (SU). See Annex 2 for details on data to be collected for WASABY aims
c) Cancer registries to have cancer rates (or estimated using data in points a and b). These data are at smallest administrative unit (SU).
d) Geographic institutes or other data sources to collect shapefiles with cancer registry area maps. See Annex 2 for details on data to be collected for WASABY aims
e) National offices or other data sources for data on deprivation indexes. These data are at smallest administrative unit (SU). See Annex 3 for details on data to be collected for WASABY aims
f) National or international agencies to reach data on environmental indicators. These data are at sample point XY coordinates. See Annex 4 for details on databases and variables identified for WASABY aims. Specific models must be performed to interpolate data from sample point coordinates to smallest administrative unit (See Annex 5).

4. ANALYSIS TO BE PERFORMED

Different type of analysis can be performed:

- Quantitative methods: generalized additive models can be used to estimate cancer risk by SU, a form of non-parametric or semi-parametric regression with the ability to analyze area-based data adjusting for covariates. The model is semi-parametric because it includes both nonparametric and parametric components. LOESS smooth is used which adapts to changes in population density where the amount of smoothing depends on the percentage of the data points in the neighborhood. This methodology is applied to estimate risk maps, adjusting for: deprivation index alone, concentration of selected pollutant/pollutants alone, deprivation index and pollutant concentration together, and with their interaction. In order to assess the contribution of these factors to the underlying spatial patterns, we should compare these maps with maps using the crude model (the unadjusted geographic variation in cancer risk). See details in Annex 6.

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 2014-2020 under Grant Agreement PP-2-5-2016 (# 769767)

- Qualitative methods: includes evaluation of sources of pollution present in the proximity of clusters identified by quantitative methods. A number of criteria must be considered for the selection of areas with an increased cancer risk, suitable for exploring the relationship between pollutants concentration and cancer. Such criteria for example include the following:
  - information on type of chemicals related with cancer risk;
  - connection between type of chemicals and different sources of pollution (e.g. industrial sites, landfills and other productive activities);
  - each target area must have an history of contamination and available data for calendar years taking into account the lag time between exposure and diagnosis (e.g. 10 years);
  - when possible, the selected polluted site is defined also in terms of its predominant surrounding conditions. In other words, the target area should be described for the vocational properties of its peripheral landscape (environments, settlements and human activities) by using general land-use indicators (urban, rural, sub-urban, natural, social, residential, industrial, etc.);
  - information on major natural (physical, geochemical and climatic) variables of the target area is recommended;
  - information on concentrations and concentration trends of pollutants over time must be available for ground waters and/or soils, and possibly expressed as TEQ (Toxic Equivalents);
  - the integrative use of data resulting from previous studies on ecological indicators, or from other investigations aimed to clarify the biological and ecological effects of chemical pollution, is not mandatory but recommended;
  - density population should be taken into consideration.

5. POSSIBLE RESULT SCENARIO

The aim of the pilot study is to perform a second-level analysis following a first-level analysis highlighting some (at least one) possible clusters of spatial incidence rate of breast cancer. The aim of the second-level analysis is to shed light on spatial clusters. The following scenarios could be given:

a) A quantitative analysis can be performed because of the availability of pollutant concentrations
b) A quantitative analysis cannot be performed because pollutant concentrations are not available: maybe we could only use pollutant emissions (not easy to be modeled) or maybe the available concentration measures are too far from the spatial cluster area to really represents the exposure of people residents near the clusters or these measures are given for a too large area.

In the a) case we can distinguish the following scenarios:

a1) The clusters previously identified are confirmed by the second-level analysis and the clusters reached the statistical significance: according to the type of pollutant source, measures to limit or to eliminate pollution sources have to be recommended. If needed an analytical study has to be planned. This last is the case if some factors might have biased the ecological, second–level analysis.

a2) The cluster identified by the first level analysis shows of the same magnitude in the second level analysis but does not reach the statistical significance: epidemiologists that performed the analysis must evaluate the hypothesis that the statistical analysis performed lacked the power to detect a real cluster. This could be done considering the relative risk indicated by published studies and the sample size of the population under study. If the hypothesis of the lack of power will be considered true and if the association between the pollutants found in water and soil and breast cancer is evaluated as strong by published studies, also in this case measures to limit or to eliminate pollution sources have to be recommended.

a3) The second level analysis does not confirm the cluster: the epidemiologists that are carrying on the study have to consider the specificity of the statistical methods used and the amount of pollution from the identified sources active on the area and to decide if perform an analytical study or to investigate other risk factors or stop analysis because no danger for the population exists.

In the b) case a decision has to be performed even if concentration pollutant information lacks: a qualitative analysis putting together the results of the first level analysis and the information about pollutants, even if scarce, must lead to a decision about population health.

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 2014-2020 under Grant Agreement PP-2-5-2016 (# 769767)

6. INT CONTACTS FOR DATA COLLECTION

Every information request and submission regarding the pilot environmental study must be addressed to:

paolo.contiero@istitutitumori.mi.it (Paolo Contiero)
alessandro.borgini@istitutotumori.mi.it (Alessandro Borgini)
martina.bertoldi@istitutotumori.mi.it (Martina Bertoldi)

and CC:

lifetable@istitutotumori.mi.it (Paolo Baili)
roberto.lillini@istitutotumori.mi.it (Roberto Lillini)

7. PUBLICATION POLICY OF THE ENTIRE PROJECT WASABY

All publications performed in the WASABY context must mention the WASABY Working Group. A suitable authorship formula being: Authors A, B, C, ... and the WASABY Working Group, with all members listed in a footnote or appendix to the article.

The WASABY Working Group will be realized with the following members:

- All members of the Steering Committee (SC)
- All members of the Management Support Team (MST)
- Up to two members of each Partner indicated in the introduction (in addition to those included in the SC and MST)
- Up to two members of each Cancer Registry participating in WASABY
- All experts actively participating in the work packages of the project
- Up to one member for each participating area working in geocoding activities (unless included in the previous points)

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 2014-2020 under Grant Agreement PP-2-5-2016 (# 769767)

ANNEX 1 – LIST OF PARTICIPATING CANCER REGISTRIES TO BE CHOSEN AFTER THE SPATIAL ANALYSIS

Country	Cancer Registry
France	Gironde
	Poitou-Charentes
	Loire-Atlantique et Vendée
	Haute Vienne
	Calvados
	Manche
	Somme
	Lille et sa région
	Bas-Rhin
	Haut-Rhin
	Doubs et Territoire de Belfort
	Cancers gynécologiques de Côte-d'Or
	Isère
	Hérault
	Tarn
	Guadeloupe
	Martinique
Germany	Schleswig-Holstein
Italy	Alto Adige
	Napoli 3 South
	Palermo
	Parma
	Ragusa
	Siracusa
	Trapani
	Trento
	Umbria
	Varese
Lithuania	Lithuania
Portugal	Central Portugal
	Northern Portugal
Slovenia	Slovenia
Spain	Basque Country
	Castellon-Valencia
	Girona
	Granada
	Murcia
	Navarra
UK	Northern Ireland
Poland	Greater Poland
	Kielce
	Silesia
	Masovia
	Podkarpackie
	Polish National Cancer Registry

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 2014-2020 under Grant Agreement PP-2-5-2016 (# 769767)

ANNEX 2 – DATA TO BE COLLECTED FOR EACH CANCER REGISTRY

FILE WITH BREAST CANCER CASES AND GEOGRAPHIC DATA

Primary invasive female breast cancer (ICD9 174*, ICD10 C50*), selected from cancer registries data during a specific ten years period (ex: 2001 to 2010) are included in the project. It is mandatory to collect data with age at diagnosis less than 50 years of age, while it is not mandatory to collect data for all ages. Synchronous and metachronous breast cancer cases must be counted once. Cancer registration criteria must follow European Network of Cancer Registries (ENCR) rules.

Residence addresses at diagnosis retrieved from the National or local Security system or from the personal data reference of each registry will be collected.

Data can be collected in two different modalities.

OPTION 1 – individual level

		Variable name	Description	Data type	Mandatory
BREAST CANCER VARIABLES		CR	Cancer Registry name	Alphanumeric variable	Yes
		PATIENT_ID	Patient identification code assigned by Cancer Registry.	Numeric / Alphanumeric variable	Yes
		DATE OF DIAGNOSIS	Incidence date based on histological or cytological confirmation of the malignancy	DD/MM/YYYY	Yes
		DATE OF BIRTH	Date of birth of the patient	DD/MM/YYYY	Yes (one of the two variables)
		AGE	Age at diagnosis	Numeric variable	
		ICDO3_M	ICDO3 morphology code of incident case	Alphanumeric variable	Yes
		SUBTYPE_ER	Estrogen Receptor value at diagnosis	Numeric / Alphanumeric variable	No
		SUBTYPE_PGR	Progesterone Receptor value at diagnosis	Numeric / Alphanumeric variable	No
		SUBTYPE_HER2	HER-2 value at diagnosis	Numeric / Alphanumeric variable	No
		SUBTYPE_KI67	KI-67 value at diagnosis	Numeric / Alphanumeric variable	No
		SUBTYPE_FISH	FISH value at diagnosis	Numeric / Alphanumeric variable	No
GEOGRAPHIC VARIABLES	OPTION A	X	Longitude coordinate referred to the address where the patient was residing at the moment of the breast cancer diagnosis	Numeric variable	Yes, data from one option
		Y	Latitude coordinate referred to the address where the patient was residing at the moment of the breast cancer diagnosis	Numeric variable	
		Reference	The coordinate system used for X and Y: UTM WGS84 32N vs. UTM ED 1950 32N	Alphanumeric variable	
	OPTION B	SU	Smallest administrative unit (SU) where the patient was residing at the moment of the breast cancer diagnosis	Alphanumeric variable	
	OPTION C	MUNICIPALITY_ CODE	Code of the Municipality where the patient was residing at the moment of the breast cancer diagnosis	Alphanumeric variable	
		MUNICIPALITY	Name of the Municipality where the patient was residing at the moment of the breast cancer diagnosis	Alphanumeric variable	

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 2014-2020 under Grant Agreement PP-2-5-2016 (# 769767)

OPTION 2 – aggregated level

		Variable name	Description	Data type	Mandatory
BREAST CANCER VARIABLES		CR	Cancer Registry name	Alphanumeric variable	Yes
		YEAR DIAGNOSIS	Incidence year based on histological or cytological confirmation of the malignancy	Numeric variable	Yes
		AGE	Age class at diagnosis	Alphanumeric variable	Yes
		ICDO3_M	ICDO3 morphology code of incident case	Alphanumeric variable	Yes
		SUBTYPE_ER	Estrogen Receptor value at diagnosis	Numeric / Alphanumeric variable	No
		SUBTYPE_PGR	Progesterone Receptor value at diagnosis	Numeric / Alphanumeric variable	No
		SUBTYPE_HER2	HER-2 value at diagnosis	Numeric / Alphanumeric variable	No
		SUBTYPE_KI67	KI-67 value at diagnosis	Numeric / Alphanumeric variable	No
		SUBTYPE_FISH	FISH value at diagnosis	Numeric / Alphanumeric variable	No
GEOGRAPHIC VARIABLES	OPTION B	SU	Smallest administrative unit (SU) where the patient was residing at the moment of the breast cancer diagnosis	Alphanumeric variable	Yes, data from one option
	OPTION C	MUNICIPALITY_ CODE	Code of the Municipality where the patient was residing at the moment of the breast cancer diagnosis	Alphanumeric variable	
		MUNICIPALITY	Name of the Municipality where the patient was residing at the moment of the breast cancer diagnosis	Alphanumeric variable	
DATA		NR_CASES	Number of primary invasive female breast cancer by all the previous variables	Numeric variable	Yes

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 2014-2020 under Grant Agreement PP-2-5-2016 (# 769767)

POPULATION FILES

For pilot environmental study, WASABY needs the reference population at the same geographic level on which that CR intends to study the incident cases. More specifically, the population files must contain the female population data by 5-year age groups, calendar year within time period and SU (sub-areas refer to the smallest geographical area for which required data are available and may be different across countries). All the variables are mandatory.

Variable name	Description	Data type
CR	Cancer Registry name	Alphanumeric variable
AGE_CLASS	5-year age class	Numeric/Alphanumeric variable
YEAR	Calendar year	Numeric/Alphanumeric variable
REF_DATE	Reference date of population data (1st Jan, 31st Dec, ecc)	Date/Alphanumeric variable
SU	MUNICIPALITY_CODE or SU indicated in the file with geographic data (see pages 5 or 6)	Alphanumeric variable
POP	Female population by 5-year age groups, calendar year within time period and sub-area on which the incidence data would be estimated	Numeric

SHAPEFILES

For the selected CR, WASABY needs a complete shapefile of the geographic area covered by its activity. The shapefile format is a digital vector storage format for storing geometric location and associated attribute information. It consists of a collection of files with a common filename prefix (e.g., Varese.shp, Varese.dbf, Varese.shx), stored in the same directory, with mandatory and optional files.

Mandatory files:

File name	Description	Data type
(CR area).shp	Shape format; the feature geometry itself	Alphanumeric
(CR area).shx	Shape index format; a positional index of the feature geometry to allow seeking forwards and backwards quickly	Alphanumeric
(CR area).dbf	Attribute format; columnar attributes for each shape, in dBase IV format	Alphanumeric

Files must be combined with information on calendar years of validity (in case of administrative changes of SU in the incidence years studied).
Other optional files, regarding spatial features not reported in the .dbf file, can be added but are not needed for a correct representation.
In the .dbf file an information about the minimum geo-coding level must be reported (i.e., census block, municipality, etc.)

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 2014-2020 under Grant Agreement PP-2-5-2016 (# 769767)

ANNEX 3 – SOCIO ECONOMIC STATUS (SES) AND OTHER CONFOUNDERS: DEPRIVATION INDEX

Since this study includes different European countries, it is important that measurement of socioeconomic deprivation be comparable or at least transferable between different European countries, despite their socio-cultural differences, to improve the comparability and reproducibility across countries. The European Deprivation Index (EDI) measures the social environment in a comparable manner across countries, despite the differences in the census variables available, and to incorporate the social and cultural specificities of each country concerned. The ecological deprivation indices are built according to shared methodological principles, by selecting fundamental needs associated with both objective and subjective poverty, and they use the same theoretical concept of relative deprivation using a European survey dedicated to relative deprivation (Eu-Silc) regularly conducted on national samples from the all European countries.

The file with confounders is structured this way:

Variable name	Description	Data type	Mandatory
COUNTRY	Country name	Alphanumeric variable	Yes
SUB_AREA	MUNICIPALITY_CODE or SU indicated in the file with geographic data (see pages 5 or 6)	Alphanumeric variable	Yes
SES_SCALE	European Deprivation Index or specific national deprivation indices (according to the availability in the specific CR) by SU of incidence data. This is a scale variable	Numeric - Scale	Yes
SES_ORDINAL	European Deprivation Index or specific national deprivation indices (according to the availability in the specific CR) by SU of incidence data, classified by deprivation groups. This is an ordinal variable	Numeric - Ordinal	Yes

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 2014-2020 under Grant Agreement PP-2-5-2016 (# 769767)

ANNEX 4 – ENVIRONMENTAL DATABASES

Databases collecting water and soil quality measures

Name	Argument	Web address	Organization	Countries* included	Years covered
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

*All the Wasaby countries (France, Germany, Italy, Lithuania, Poland, Portugal, Spain, Northern Ireland, Slovenia) are covered by the different databases.

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 2014-2020 under Grant Agreement PP-2-5-2016 (# 769767)

Available data regarding the main organic and inorganic persistent contaminants that we considered as prior in the relationship with breast cancer through articles visible in the following link

Pollutants	Unit of measurement	Database	Environmental Matrix	Measure periodicity	Spatial resolution
PCBs pcb total (congeners 28, 52, 101, 118, 138, 153, 180)	µg/l, point concentration	Waterbase groundwater **I**PCHEM	Water	annual	Sample points
DDT, DDD, DDE	µg/l, point concentration	Waterbase groundwater **I**PCHEM	Water	annual	Sample points
DDT	Kg/year	E-PRTR industrial point emissions	in Air, Water and Land threshold for releases 1 kg/ year	annual	Emission points
Alachlor	Kg/year	E-PRTR industrial point emissions	Water and Land threshold for releases 1 kg/ year	annual	Emission points
Aldrin	Kg/year	E-PRTR industrial point emissions	in Air, Water and Land threshold for releases 1 kg/ year	annual	Emission points
Gamma-hch (Lindane)	µg/l, point concentration	Waterbase groundwater **I**PCHEM	Water	annual	Sample points
	Kg/year	E-PRTR industrial point emissions	in Air, Water and Land threshold for releases 1 kg/ year	annual	Emission points
Hexachlorobenzene (hcb)	µg/l, point concentration	Waterbase groundwater **I**PCHEM	Water	annual	Sample points
	Kg/year	E-PRTR industrial point emissions	in Air threshold for releases 10 kg/ year	annual	Emission points
			in Water and in land threshold for releases 1 kg/ year	annual	Emission points
Chlordane	µg/l, point concentration	Waterbase groundwater **I**PCHEM	Water	annual	Sample points
PAH benzo(a)pyrene benzo(a)anthracene	µg/l, point concentration	Waterbase groundwater **I**PCHEM	Water	annual	Sample points
Anthracene	Kg/year	E-PRTR industrial point emissions	in Air threshold for releases 50 kg/ year	annual	Emission points
			in Water threshold for releases 1 kg/ year	annual	Emission points
			in Land threshold for releases 1 kg/ year	annual	Emission points

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 2014-2020 under Grant Agreement PP-2-5-2016 (# 769767)

Pollutants	Unit of measurement	Database	Environmental Matrix	Measure periodicity	Spatial resolution
Triazine atrazine desethylatrazine desisopropylatrazine	µg/l, point concentration	Waterbase groundwater IPCHEM	Water	annual	Sample points
Atrazine	Kg/year	E-PRTR industrial point emissions	in Water threshold for releases 5 kg/ year	annual	Emission points
		E-PRTR industrial point emissions	in Land threshold for releases 5 kg/ year	annual	Emission points
Cadmium	µg/l, point concentration	Waterbase groundwater IPCHEM	Water	annual	Sample points
	mg/kg, area concentration	Geochemical Atlas of Europe	Soil	2008	4.700 km²
	mg/kg, area concentration	LUCAS topsoil database	Soil	2009 and 2012	200 km² for one sample
	Kg/year	E-PRTR industrial point emissions	in Air threshold for releases 10 kg/ year	annual	Emission points
		E-PRTR industrial point emissions	in Water threshold for releases 5 kg/ year	annual	Emission points
		E-PRTR industrial point emissions	in Land threshold for releases 5 kg/ year	annual	Emission points
PCDD (dioxins	Kg/year	E-PRTR industrial point emissions	in Air, Water and land threshold for releases 0,0001 kg/ year	annual	Emission points
Trihalomethanes (THMs) bromoform bromo-dichloro-methane dibromochlorometh ane chloroform	µg/l, point concentration	Waterbase groundwater IPCHEM	Water	annual	Sample points

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 2014-2020 under Grant Agreement PP-2-5-2016 (# 769767)

ANNEX 5 – INTERPOLATION METHODS

Interpolation of the pollutant data from the source to the considered area of analysis (SU)

If smaller geographic unit is considered for collecting information about pollution with respect to health outcomes, different strategies will be considered and evaluated in order to work on the misalignment of the data, as found in the WASABY review submitted to the "Reviews of environmental contamination and toxicology":

- The Kriging interpolation method (a Gaussian process regression) to extend the information collected at the pollution sources to the areas considered for epidemiologic population data (Colak et al. 2015, Lin et al. 2014, López-Abente et al. 2018, Núñez et al. 2016, Núñez et al. 2017).
- A combination of Ordinary Least Squares (OLS) regression models and Geographical Weighted Regression (GWR) models, corrected by spatial lag models, after testing the presence of local spatial autocorrelation by Local Indicators of Spatial Association (Hanchette et al. 2018).
- A two-level model to estimate the effects of the pollution in sampling points to the whole districts, with pollution included as a random intercept and correcting the misalignment by adaptive quadrature method (Aballay et al., 2012).

A preliminary exploratory spatial analysis using Moran's I for testing spatial autocorrelation after data geo-coding procedures will be considered.

Kriging interpolation algorithm is implemented in R and other free software; Moran's I and the other procedures are implemented both in Stata and R software.

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 2014-2020 under Grant Agreement PP-2-5-2016 (# 769767)

ANNEX 6 – REGRESSION MODELS

Models to be used for the spatial analysis of the pilot environmental study.

A first considered approach is an inferential one: the Spatial Autoregressive with Autoregressive Disturbance model (SARAR) (see deliverable "D4.3 - Cancer incidence and water & soil pollutants: Data management model to perform spatial analysis").

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

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

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, both environmental and socio-economic), 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).

Spatial interactions are modeled 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.

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

SARAR models are implemented in Stata and R software.

Geographic representation of the results of pilot environmental study.

Software to be used for the above procedures GIS software such as QGis [http://www.qgis.org/en/site/] and ArcGis desktop 10.0 will be used to create maps with many layers (raster or vector) using different map projections.

The vector data are stored as either point, line, or polygon-feature. Different kinds of raster images are supported, and the software can geo-reference images. Maps can be assembled in different formats and for different uses.

They also could be used to improve, if necessary, the spatial 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 2014-2020 under Grant Agreement PP-2-5-2016 (# 769767)

ANNEX 7 – REFERENCES

Aballay LR, Díaz Mdel P, Francisca FM, Muñoz SE. 2012. Cancer incidence and pattern of arsenic concentration in drinking water wells in Córdoba, Argentina. Int J Environ Health Res 22(3):220-231.

Colak EH, Yomralioglu T, Nisanci R, Yildirim V, Duran C. 2015. Geostatistical analysis of the relationship between heavy metals in drinking water and cancer incidence in residential areas in the Black Sea region of Turkey. J Environ Health 77(6):86-93.

Hanchette C, Zhang CH, Schwartz GG. 2018. Ovarian Cancer Incidence in the U.S. and Toxic Emissions from Pulp and Paper Plants: A Geospatial Analysis. Int J Environ Res Public Health 15(8).

Lin WC, Lin YP, Wang YC, Chang TK, Chiang LC. 2014. Assessing and mapping spatial associations among oral cancer mortality rates, concentrations of heavy metals in soil, and land use types based on multiple scale data. Int J Environ Res Public Health 11(2):2148-2168.

López-Abente G, Locutura-Rupérez J, Fernández-Navarro P, Martín-Méndez I, Bel-Lan A, Núñez O. 2018. Compositional analysis of topsoil metals and its associations with cancer mortality using spatial misaligned data. Environ Geochem Health 40(1):283-294.

Núñez O, Fernández-Navarro P, Martín-Méndez I, Bel-Lan A, Locutura JF, López-Abente G. 2016. Arsenic and chromium topsoil levels and cancer mortality in Spain. Environ Sci Pollut Res Int. 23(17):17664-17675.

Núñez O, Fernández-Navarro P, Martín-Méndez I, Bel-Lan A, Locutura Rupérez JF, López-Abente G. 2017. Association between heavy metal and metalloid levels in topsoil and cancer mortality in Spain. Environ Sci Pollut Res Int. 24(8):7413-7421.

WASABY Deliverables (available at http://www.wasabysite.it/):

- D4.3 - Cancer incidence and water & soil pollutants: Data management model to perform spatial analysis
- D5.1 - Deprivation index available data
- D6.1 - Report on methods for spatial analysis of the cancer data
- D7.1 - Literature review on the main persistent environmental contaminants related to breast cancer
- D7.2 - Report on environmental data available for spatial 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 2014-2020 under Grant Agreement PP-2-5-2016 (# 769767)