International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

Classification-driven air pollution mapping as for environment and health analysis

S. Wiemann [a], **S. Richter**[a], **P. Karrasch**[a], **J. Brauner**[a], **K. Pech**[b], **L. Bernard**[a]

[a] *Professorship of Geoinformation Systems, Technische Universität Dresden*
(Stefan.Wiemann@tu-dresden.de, Silke.Richter@tu-dresden.de,
Pierre.Karrasch@tu-dresden.de, Johannes.Brauner@tu-dresden.de,
Lars.Bernard@tu-dresden.de)
[b] *Chair of Photogrammetry, Technische Universität Dresden*
(Katharina.Pech@tu-dresden.de)

Abstract: This article presents an air pollution modelling approach and its use in health applications within the EO2HEAVEN project. The model is based on a multidimensional Inverse Distance Weighting and makes use of artificial distances between area attributes. The number and type of attributes used is fully customizable and can be adapted according to specific application fields and data preconditions. It is kept flexible and simple and thus, suitable to be used within a Spatial Data Infrastructure to provide access to real-time air pollution information via the internet. In a prototypical implementation the model is applied to estimate the concentration of particular matter and ozone in the Federal State of Saxony, Germany.

Keywords: Air pollution; Modelling; Health Applications; Spatial Data Infrastructures

1 INTRODUCTION

The adverse effect of air pollution to human health is already well known and documented (Janssen and Mehta [2006], EEA [2011]). Various studies have shown that acute as well as chronic respiratory and cardiovascular diseases can be directly linked to high air pollution concentrations. In many countries, legislative initiatives helped to reduce air pollution within the past decades. However, air pollution is still considered a significant thread to human health (WHO [2009]).

Assessing a personal health risk from air pollution usually requires information on personal exposure and dose (Özkaynak [1999]). However, as both are difficult to measure for larger population groups, air pollution concentration is usually used as a proxy. For the European Union, the Directive 2008/50/EC defines regulations and requirements for corresponding air quality measurements (EU [2008]) and thus, lays the framework for European health risk assessment from air pollution.

The modelling of continuous air pollution information can be considered as an important task for environmental health risk analysis and prevention. Today, various in situ sensor networks, either governmental, commercially or voluntarily driven, can provide real time information on air pollution. In addition, a number of numerical models have been applied to derive corresponding spatio-temporal distributions. These models offer valuable results but are also quite demanding, for instance in terms of input variables, user skills and computing power. Thus, a need for robust and less demanding methods is emerging especially for real-time applications, such as an online system in a Spatial Data Infrastructure (SDI). SDIs allow for an up-to-date and interoperable access to processed data and relevant data sources (Groot and McLaughlin [2000]). Automatically collected air pollution measurements can be provided over the internet to enable online processing

and result visualization. For instance, the EO2HEAVEN (Earth Observation and ENVironmental modelling for the mitigation of HEAlth risks) project is developing an online system using these SDI concepts.

2 State of the art in air quality modelling for health risk assessment

Based on Johnson et al. [1997], air quality can be defined as a measurement of the condition of air with respect to human needs. The process for describing and quantifying the health risk is considered as health risk assessment and typically consists of four steps: hazard identification, dose-response assessment, exposure assessment, and risk characterization (NRC [1983]). With respect to the "environmental pathway", which follows the source, emission, concentration, dose and health effect of air pollutants (Janssen and Mehta [2006]), health risk assessment related to air quality essentially relies on air pollution modelling and corresponding health impact studies.

A brief summary on the history of air pollution modelling, starting with the modelling of industrial plumes in the early 1930s, is given by Daly and Zannetti [2007]. To date, various air pollution models have been developed for different regions, scales and pollutants. An overview on commonly used global and European regional models is reviewed by Huijnen et al. [2010]. Moreover, various air pollution and dispersion models applicable from local to global geographic scales are summarized by the COST[1] project model inventory.

Due to different input data, assumptions and modelling strategies, each air pollution model leads to different results. Validation is performed either by statistical analysis on a subset of input data (see Janssen et al. [2008], Slørdal et al. [2008], Beelen et al. [2009]) or by comparison with external datasets usually applying statistical methods (see Engel-Cox et al. [2004], Petrakis et al. [2005]). However, while the former strongly depends on the selected subset and the latter is hampered by the missing "gold standard" for air pollution modelling, an ensemble of models might be applicable to predict air quality in a probabilistic manner (Huijnen et al. [2010]).

To obtain or estimate air pollution information, there are three commonly used approaches, which can be used in any combination: (1) emission modelling from pollution sources including dispersion modelling, (2) interpolation of pollutant measurements and (3) estimation of pollutant concentrations based on natural and anthropogenic characteristics. The first approach relies on emission information for pollutant sources like industrial complexes, traffic or combustion heating. However, exact emission rates and the proportion between natural and anthropogenic sources is often uncertain as indicated by Klingner and Sähn [2008] (particular matter), Vingarzan [2004] (ozone) and Olivier et al. [1990] (nitrogen oxides). The second approach uses discrete or continuous air pollutant measurements from in situ or remote sensor systems and derives coverage information by using geographical interpolation methods (EPA [2009], Pebesma et al. [2011]). As the quality of interpolation depends essentially on the number of observations, it can be complemented with the application of virtual sensors as presented by Ranchin and Wald [2010]. The third approach to estimate air pollution is based on the natural and anthropogenic characteristics of the investigated area and assumes that air pollution mainly depends on local characteristics, such as land cover, surface structure, population density or traffic density. It has already been applied by various research studies (see Umweltbundesamt [2005], Janssen et al. [2008] or Beelen et al. [2009]) and is referred to as reliable for studies on long-term exposure (Gulliver et al. [2011]).

3 An approach for real time air pollution modelling

The proposed real-time air pollution modelling approach in this paper is based on Toblers 1st law of geography, which states that "Everything is related to everything else, but near

[1] http://www.mi.uni-hamburg.de/costmodin

things are more related than distant things" (Tobler [1970]). However, distance here is not exclusively seen as a geographic distance, but as a measurement of similarity between two distinct areas, called attribute distance. The assumption is that areas with a close attribute distance are more likely to share similar air pollution characteristics than more distant ones.

McGregor [1996] identified four distinct affinity areas for SO_2 pollution in the area of Birmingham using principal component and cluster analysis based on a network of 17 in situ stations. Within the APMoSPHERE project, a similar approach has been applied on the European scale. Therein, a clustering of land cover, road network, meteorological and topographic characteristics led to a subdivision into 10 up to 14 affinity zones (APMoSPHERE [2005]). However, certain drawbacks for standard clustering methods need to be considered. Many clustering algorithms are quite sensitive to input parameters such as the chosen number of clusters or the initial seed point. The clustering results are often difficult to interpret, especially when dealing with high dimensions (Han and Kamber [2006]). Furthermore, the explicit cluster boundaries implied by most cluster algorithms might not sufficiently represent the continuous landscape. In contrast to previous studies, we do not calculate distinctive clusters, but attribute distances for each area unit to existing in situ stations to express how far the area can be represented by each single station. Similar to the method of Inverse Distance Weighting (IDW) by Shepard [1968], the total attribute distance is depending on the inverse distance between the selected attributes as follows:

$$v(a) = \sum_{i=0}^{N} \frac{\omega_i(a) v_i}{\sum_{j=0}^{N} \omega_j(a)}, \qquad N = \text{number of observations} \tag{1}$$

The estimated value $v(a)$ is calculated over the sum of specifically weighted observation measurements v_i. As the proposed method may include an arbitrary number of attributes a, the distance weight $\omega(a)$ can be expressed as:

$$\omega(a) = \sum_{i=0}^{M} \frac{\gamma_i \frac{1}{d(a,a_i)^{p_i}}}{\sum_{j=0}^{M} \gamma_j}, \qquad M = \text{number of attributes} \tag{2}$$

Thus, the specific weight is calculated over the sum of weighted inverse distances $d(a, a_i)$ between the selected attributes. The specific weight γ depends on single pollutant characteristics and must be subject to prior statistical analysis of the available observations. As for the standard IDW method, the distance weight power p_i is used to better reflect the decrease of influence of a certain parameter with an increasing distance to the origin. To ensure comparability, the single attribute distances have to be normalized by dividing the absolute attribute distance by the maximal attribute value.

To calculate a distance between observations and single reference areas, a suitable area around the corresponding in situ station location needs to be defined and classified as well. The greatest challenge here is to find the most suitable area to represent an observation. In a comparable Belgian study by Janssen et al. [2008], a radius of 2km around each station was found to be most appropriate. We generalize this assumption and propose that the station classification for each attribute should be calculated based on a weighting of classifications using different radiuses in the following way:

$$a(r) = \sum_{i=0}^{N} \frac{\omega_i \frac{a_i}{2\pi r_i^2}}{\sum_{j=0}^{N} \omega_j}, \qquad N = \text{number of radiuses} \tag{3}$$

An attribute $a(r)$ is calculated over a weighted sum of attribute appearance a_i within surrounding areas of varying radiuses r_i. The specific weight ω_i is dependent on single

attribute characteristics and the type of station and should be subject to prior analysis.

The estimation of air pollutant concentrations can be based on either the whole set of observations or only those which are closest concerning their attribute distance. To achieve most reliable results, the chosen attributes should reflect the influences on pollutant concentrations as completely as possible. If only static input data is used for attribute distance calculation, the model is suitable to be used in real-time applications, because of the fact that the distances between each station and the reference areas must be calculated only once. However, as the model is only capable of giving a rough estimation of air pollution concentrations in real-time, it should be complemented with existing tools to produce refined air pollutant concentrations for political decision making or archiving purposes. Especially the inclusion of real-time pollutant emission and dispersion modelling is expected to greatly improve the results.

4 IMPLEMENTATION AND EVALUATION OF THE MODEL

A prototype of the model, implemented in R, is applied for particular matter (PM_{10}) and Ozone (O_3) for the Federal State of Saxony, Germany in the context of the EO2HEAVEN project. Therefore, the official sensor network operated by the Saxon State Office for Environment, Agriculture and Geology (LfULG) is used, providing 26 stations measuring PM_{10} and 24 stations measuring O_3 concentrations. Because of the small number of observations, standard interpolation techniques are difficult to apply. Thus, the previously presented method is used to calculate area wide pollutant concentrations.

To realize the proposed method for attribute distance calculation, the following attributes are chosen and mapped to a 1x1km raster for the area of Saxony:

- Land cover information taken from the Corine Land Cover (CLC) dataset from 2006 at a 100m resolution. To reduce the number of attributes, a land use indicator is calculated for each pollutant in the same manner as described by Janssen et al. [2008]
- Elevation data from the Shuttle Radar Topography Mission (SRTM) at a 90m resolution
- Population density data obtained from the European Environment Agency (EEA) and corrected using official statistics on the municipality level
- Road density information from OpenStreetMap (OSM) calculated from the main street feature classes in OSM in road kilometre per square kilometre.
- Traffic census data from LISt GmbH for the major roads in Saxony in vehicles per square kilometre

Based on the selected attributes, PM_{10} and O_3 maps have been calculated for the daily average observations for the years 2003 to 2007 and for half hourly observations for the year 2006. The deduced five year average maps from 2003 to 2007 are depicted in Figure 1. A selection of corresponding distance maps is depicted in Figure 2 indicating how well a certain area is represented by the station observations.

The weights for the chosen attributes are obtained from a maximization of the correlation between the PM_{10} and O_3 measurement correlation and the calculated attribute difference between each station. The chosen weights and corresponding correlations used for the implementation are listed in Table 1. Respective scatterplots are depicted in Figure 3. Although the correlations, especially for PM_{10}, are quite low and the weights must not necessarily reflect causal interrelations, they are considered as applicable with respect to the current implementation.

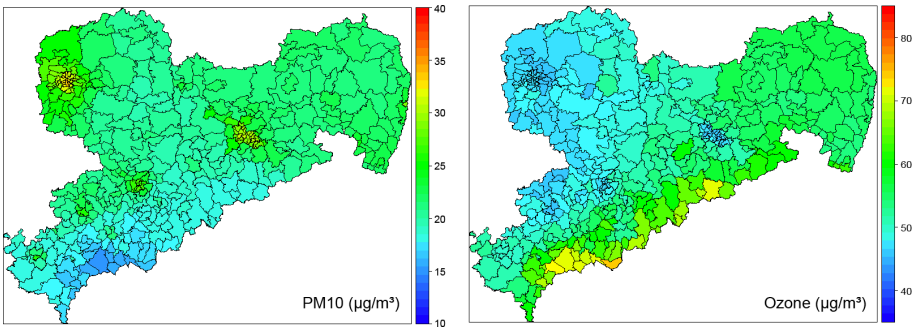

Figure 1: Four year average concentrations for PM_{10} (left) and O_3 (right) as estimated by the model, aggregated on postal code level; blue = low concentration, red = high concentration

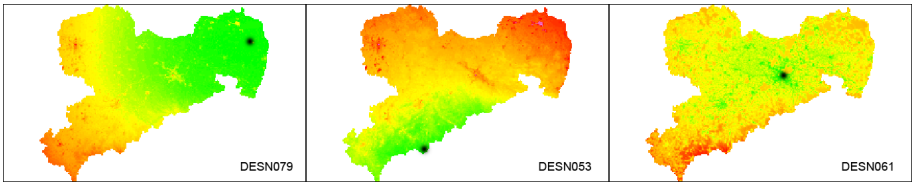

Figure 2: Attribute distance index for selected in situ stations; green = area is well represented, red = area is not represented

	LC	EL	PD	RD	TC	GD	R^2
O_3	30	20	0	5	15	30	*0.62*
PM_{10}	30	10	0	15	15	30	*0.35*

Table 1: Attribute weights and corresponding coefficient of determination R^2 between measurement correlations and attribute distances between stations; LC = Land Cover, EL = Elevation, PD = Population Density, RD = Road Density, TC = Traffic Census, GD = Geographic Distance

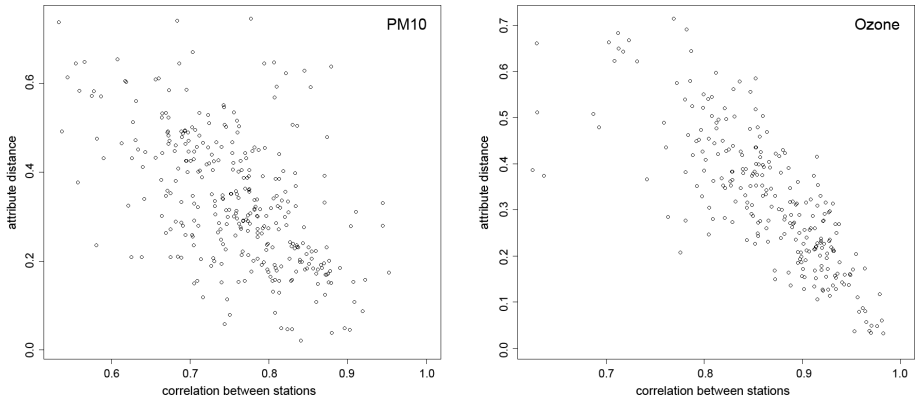

Figure 3: Scatterplots for the correlation of measurements (x-axis) and the calculated attribute distance (y-axis) for PM_{10} and O_3

To give estimates on the model accuracy and uncertainty, internal and external validation methods have been applied on the model results, in particular:

- Internal cross-validation and model stability tests including a random removing of observations from the model and comparison of the results. As stations vary in their significance to the model, the number and type of the removed stations considerably influences the results. Nevertheless, the variance in cross-validation did not exceed 5µg/m^3 for both PM_{10} and O_3 in the 4 year average.
- External validation of the 5-year average with an official air quality model provided by the LfULG, which is available as a 4-year average from 2001-2005. The R^2 are 0,64 for O_3 and 0,86 for PM_{10}. However, as both are only estimations, the expressiveness of this test is quite limited.

For the Saxony test case the pre-processing of attribute distances between each of the in situ stations and the 1x1km raster cells takes about one minute. The subsequent computation of an air pollution map, including the transfer of sensor measurement data across the network, takes only a few seconds. As the pre-processing is required only once, the application can be considered as suitable for a real time internet application.

5 USE OF THE MODEL FOR HEALTH DATA APPLICATIONS

In the context of the EO2HEAVEN project, the modelled environmental data for PM_{10} and O_3 concentrations is used for a correlation analysis with health data, obtained from the Allgemeine Ortskrankenkasse (AOK), one of the largest public health insurance companies in Saxony. The aim is to assess the impact of both pollutants on human health, especially the cardiovascular and respiratory systems. Because of data protection issues, health data is aggregated on postal code areas. To link the environmental and health data, the pollutant concentrations are aggregated accordingly.

People suffering from pneumonia in 2007 have been chosen to examine the influence of PM_{10} and O_3 concentrations on the respiratory system. Within our study no spatial correlations can be identified for the level of pollutant concentration and number of pneumonia patients. The same applies for medical treatments on asthma bronchiale, bronchitis and other respiratory diseases, even if a time lag between pollutant concentrations and medical treatments is considered. However, this does not mean that there is no connection between pollutant concentrations and the mentioned diseases. It just points to a much more complex interaction between them. Thus, further analysis reflecting additional influences, such as socio-economic structures, smoking or personal lifestyle will be conducted.

The results presented in this paper do only cover people insured at the AOK. Although it is one of the largest health insurance companies in Saxony, there is a bias in the health dataset due to the fact that a disproportionately high number of elderly people are insured. In order to test whether the results are valid for the whole population of Saxony, the same analysis will be performed on official morbidity and mortality statistics. Furthermore, the influence of the first and second derivation of the modelled pollutant concentrations, representing the temporal change of concentrations and the corresponding speed of change, will be tested.

6 FURTHER DEVELOPMENTS

In order to improve the presented implementation of the model, the application of real-time emission and meteorological data will play a major role. Officially reported emission data is usually only available in low spatial and temporal resolution. Thus, advanced disaggregation methods are required to achieve as accurate as possible real-time emission information. Meteorological data can be used for dispersion modelling in order to add

forecasting functionality to the model. Additional internal and external validation methods, either statistically or empirically driven, need to be applied to express and refine applicability and uncertainty information. Therefore, the use of remote sensing data is envisaged.

Technical challenges need to be addressed if the presented methods are used in real-time in an SDI. Although the presented approach does normally not involve tedious and long-time calculations used in numerical models, it is still a time consuming process to calculate results and provide them in real-time as a map. Concepts from High Performance Computing can be considered to speed up the calculations.

The acceptance of the results among public health practitioners or politicians for instance, is of high importance. A transparent workflow and appropriate metadata for the distributed source data sets and processing steps are needed to provide trustworthy results as a base for political decision making and further developments. Means to transport the data lineage and provenance information throughout the whole workflow are required. The robustness of the model needs to be evaluated with respect to different in situ station settings, input parameters and attribute weights. The transferability to different regions is required to facilitate the use of the model for health applications.

As the work in EO2HEAVEN is also applied in Durban, Southern Africa, sufficient network bandwidth and internet access cannot be assured. Thus, offline solutions and mobile phone applications are required as well.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 244100.

REFERENCES

APMoSPHERE. 6. detailed report, related to overall project duration. Technical report, APMoSPHERE project consortium, 2005.

Beelen, R., G. Hoek, E. Pebesma, D. Vienneau, K. de Hoogh, and D. J. Briggs. Mapping of background air pollution at a fine spatial scale across the european union. *Science of The Total Environment*, 407(6):1852–1867, 2009.

Daly, A. and P. Zannetti. *Air Pollution Modeling - An Overview*, chapter 2. The Arab School for Science and Technology and The EnviroComp Institute, 2007.

EEA. Air quality in europe. Technical Report 12/2011, European Environment Agency, 2011.

Engel-Cox, J., C. Holloman, B. Coutant, and R. Hoff. Qualitative and quantitative evaluation of modis satellite sensor data for regional and urban scale air quality. *Atmospheric Environment*, 38(16):2495–2509, 2004.

EPA. Technical assistance document for the reporting of daily air quality - the air quality index (aqi). Technical report, U.S. Environmental Protection Agency, 2009.

EU. Directive 2008/50/EC of the European Parliament and of the Council of 21 may 2008 on ambient air quality and cleaner air for Europe, 2008.

Groot, R. and J. McLaughlin. *Geospatial data infrastructure: concepts, cases, and good practice*. Oxford University Press, 2000.

Gulliver, J., K. de Hoogh, D. Fecht, D. Vienneau, and D. Briggs. Comparative assessment of gis-based methods and metrics for estimating long-term exposures to air pollution. *Atmospheric Environment*, 45:7072–7080, 2011.

Han, J. and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2 edition, 2006.

Huijnen, V., H. J. Eskes, A. Poupkou, H. Elbern, K. F. Boersma, G. Foret, M. Sofiev,

A. Valdebenito, J. Flemming, O. Stein, A. Gross, L. Robertson, M. D'Isidoro, I. Kioutsioukis, E. Friese, B. Amstrup, R. Bergstrom, A. Strunk, J. Vira, D. Zyryanov, A. Maurizi, D. Melas, V.-H. Peuch, and C. Zerefos. Comparison of omi no_2 tropospheric columns with an ensemble of global and european regional air quality models. *Atmospheric Chemistry and Physics*, 10(7):3273–3296, 2010.

Janssen, N. and S. Mehta. *Human exposure to air pollution*, chapter 3. World Health Organization, 2006.

Janssen, S., G. Dumont, F. Fierens, and C. Mensink. Spatial interpolation of air pollution measurements using corine land cover data. *Atmospheric Environment*, 42(20):4884–4903, 2008.

Johnson, D., S. Ambrose, T. Bassett, M. Bowen, D. Crummey, J. Isaacson, D. Johnson, P. Lamb, M. Saul, and A. Winter-Nelson. Meanings of environmental terms. *Journal of Environmental Quality*, 26(3):581–589, 1997.

Klingner, M. and E. Sähn. Prediction of PM_{10} concentration on the basis of high resolution weather forecasting. *Meteorologische Zeitschrift*, 17(3):263–272, June 2008.

McGregor, G. Identification of air quality affinity areas in Birmingham, UK. *Applied Geography*, 16(2):109–122, 1996.

NRC. *Risk Assessment in the Federal Government: Managing the Process*. National Research Council (U.S.). Committee on the Institutional Means for Assessment of Risks to Public Health, 1983.

Olivier, J., A. Bouwman, K. Van der Hoek, and J. Berdowski. Global air emission inventories for anthropogenic sources of NO_x, NH_3 and N_2O in 1990. *Environmental Pollution*, 102(1):135–148, 1990.

Özkaynak, H. Exposure assessment. In *Air Pollution and Health*, chapter 9. Academic Press, 1999.

Pebesma, E., D. Cornford, G. Dubois, G. Heuvelink, D. Hristopulos, J. Pilz, U. Stöhlker, G. Morin, and J. Skøien. Intamap: The design and implementation of an interoperable automated interpolation web service. *Computers & Geosciences*, 37(3):343–352, 2011.

Petrakis, M., T. Kopania, B. Psiloglou, D. Briggs, G. Hoek, A. Aaheim, G. Shaddick, N. Sifakis, and A. Retalis. Gis and remote sensing techniques in emission mapping for health management in europe. *IASME Transactions*, 2:383–388, 2005.

Ranchin, T. and L. Wald. Data fusion in remote sensing of urban and suburban areas. In Rashed, T. und Jürgens, C., editor, *Remote Sensing of Urban and Suburban Areas*, volume 10 of *Remote Sensing and Digital Image Processing*, chapter 11. Springer, 2010.

Shepard, D. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, ACM '68, pages 517–524, New York, NY, USA, 1968. ACM.

Slørdal, L., H. Mc Innes, and T. Krognes. The air quality information system airquis. *Information Technologies in Environmental Engineering*, 1:40–47, 2008.

Tobler, W. A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46(2):234–240, 1970.

Umweltbundesamt. Abschätzung der Gesundheitsauswirkungen von Schwebestaub in Österreich. Technical report, Umweltbundesamt GmbH, Wien, 2005.

Vingarzan, R. A review of surface ozone background levels and trends. *Atmospheric Environment*, 38(21):3431–3442, 2004.

WHO. Global health risks - mortality and burden of disease attributable to selected major risks. Technical report, World Health Organization, 2009.