REMOTE SENSING AND HEALTH DATA FUSION: METHODOLOGICAL CHALLENGES IN CHOLERA RESEARCH

SIMONIS INGO[1], VAN DER MERWE MARNA[2], VAHED ANWAR[2]

[1]Open Geospatial Consortium Europe, London, UK - isimonis@opengeospatial.org
[2]Council for Industrial and Scientific Research, Pretoria, South Africa - (MvdMerwe2, avahed@csir.co.za

Abstract - The use of remote sensing data and technologies over the past thirty years has advanced research in spatial epidemiology. Remote sensing products often show significant correlations with cholera epidemics on a seasonal scale in various areas of the world. This paper focuses on the methodology for the use and fusion of remote sensing data for inland scenarios such as Lake Victoria. It addresses the complexities when analysing earth observation data together with health data. Contributing factors include data uncertainties and data collection and reporting protocols which are often influenced by a historically induced early institutional separation of infectious disease control (health institutions) and environmentally caused toxicological research. The proposed methodology highlights these aspects, describes approaches to support the efficient integration of in-situ sensors placed in and around inland water bodies, and reflect on new requirements related to remote sensing products and workflow tools.

Keywords: Earth Observation, Health, Cholera, data fusion, research methodology

1. INTRODUCTION

The use of remote sensing data and technologies over the past thirty years has advanced research in spatial epidemiology. Remote sensing products, in particular sea surface temperature and height, and chlorophyll-a show significant correlations with cholera epidemics on a seasonal scale in various areas of the world. Most studies however focus on the use of ocean colour datasets for coastal and estuarine areas. Moving inland towards great fresh water lakes such as Lake Victoria in Uganda and/or equatorial areas presents new challenges to epidemiologists. Typical ocean colour products cannot be used directly, as atmospheric correction usually involves a clear distinction between sea and land, differences in altitude and the influence of cloud cover at a very early step of data processing and correction. Etiological cause-effect hypotheses need to be revaluated and research methodologies applied for ocean and coastal areas need to be adapted.

This paper focuses on the methodology for the use and fusion of remote sensing data for inland scenarios such as Lake Victoria, following the data fusion definition given by [OGC2010], which reads "the act or process of combining or associating data or information regarding one or more entities considered in an explicit or implicit knowledge framework to improve one's capability (or provide a new capability) for detection, identification, or characterization of that entity". The paper addresses the complexities when analysing earth observation data together with health data. Contributing factors include data uncertainties and data collection and reporting protocols which are often influenced by a historically induced early institutional separation of infectious disease control (health institutions) and environmentally caused toxicological research. The proposed methodology highlights these aspects, describes approaches to support the efficient integration of in-situ sensors placed in and around inland water bodies, and reflect on new requirements related to remote sensing products and workflow tools.

2. CHOLERA DYNAMICS AND COMPLEXITY

Cholera case data follows a strong seasonal pattern depending on the spatial location of the area. The seasonal pattern becomes less distinct closer to the equator. The relative importance of different environmental factors driving cholera outbreaks and patterns within the outbreak data and the presence of bacteria in water sources change over space and time. Cholera outbreaks in areas where the bacteria are known to be endemic do not necessarily occur even when the environmental conditions are favourable. It is therefore often difficult to quantify the risk of an outbreak due to unexpected scenarios.

The occurrence, spread, and extent of a cholera outbreak are extremely difficult to predict and, when possible, requires the analysis of each possible situation. It is difficult to develop a generalised prediction and/or simulation model as the conditions and the relationships between environmental and case data change over time and space. Depending on the research question or aim of an analysis of environmental data together with health data, the environmental data from for example satellite and/or in situ sensors may not necessarily overlap in time and space with the health data. For example when investigating a potential link between ENSO (El Niño-Southern Oscillation) and cholera outbreaks in an area, sea surface temperature in regions far away from the area of interest may be used in the analysis. Rainfall occurring upstream from a cholera reporting area next to a river may be more important than the rainfall in the actual area of interest. The cumulative rainfall value for a period of time beforehand has been shown to be more important than the rainfall value on the day when cases are reported [Woodborne2008].

The goal of environmental data and health data fusion is to understand the relationships between occurrence and state of the Vibrio cholerae bacterium, cholera disease outbreaks, and environmental parameters. The relative importance of different factors may change due to spatiotemporal adjacency, and the scales at which the environmental, climate and social factors operate and affect disease outbreak patterns over time and space. To ground the methodologies developed later in this paper, we will briefly outline the research questions addressed in epidemiologic and environmentally caused toxicological cholera research. This is necessary as it will assist the selection of appropriate datasets and formats in terms of temporal and spatial resolution and coverage, analysis tools and presentation tools, e.g. maps, graphs etc.

2.1 Thematic Relationships

Research in spatial epidemiology has analysed potential correlations between different environmental parameters and mainly outbreaks of the cholera disease [Rodó2002;

DeMagny2008] and it seems that the type and nature of verified correlation vary across locations, seasons, and years. The research on the dynamics of the cholera bacterium is ongoing and new discoveries are still made [Zo2008], but needs to be addressed here either way to ensure a robust methodology for successful data fusion. Thus, the following questions need to be answered on a per case basis:

- What are the most important environmental (including climatic) variables driving V. cholerae dynamics in a given water body?
- What drives the changes in the relevant environmental variables, i.e. physiochemical and ecological and biological features?
- What or which environmental factors potentially drive the start of an outbreak in a given area (where area refers to a city or town/district/municipality/country/region)?

2.2 Spatial Relationships

In addition to the thematic relationships, the spatial variability as well as extent of both the bacteria and the outbreak are of major interest. In some cases, it is even the combined spatial and temporal nature of the relationship, as shown at the end of the list of cholera research questions:

- What drives the variability in terms of the presence of Vibrio cholerae in different water bodies during an outbreak in the defined area or between different sampling points within a given water body at any point in time?
- What drives the variability in the location of the first cases reported during outbreaks and as an outbreak progresses, between different areas within the defined area?
- What drives the spread of an outbreak within the defined area or between defined areas?
- At what scale do the relevant environmental parameters operate and affect the local health conditions?
- How does the relative importance of a specific environmental variable (including climatic) or combination of variables change over time in terms of disease and pathogen dynamics between different areas of interest or within a defined area?

2.3 Temporal Aspects of Relationships

The temporal aspects including variability, scale, and time lags conclude the list of general research questions:

- What temporal patterns are present in the time series of individual environmental and health variables?
- How do the relative importance of an important environmental variable (including climatic) or combination of important variables change over time and space?
- What or which environmental factors potentially drive the start of an outbreak in a given area (where area refers to a city or town/district/municipality/country/region)?
- What or which environmental factors potentially drive the duration of an outbreak in the defined area?
- What or which environmental factors potentially drive the end of an outbreak in the defined area?
- Do observed patterns in the time series data for individual environmental variables consistently occur at a specific frequency or period of time before an outbreak starts and/or ends?
- Do observed patterns in case data consistently lag patterns observed in the environmental data with a fixed amount or period of time?

3. DATA FUSION CHALLENGES

In general, the following challenges or problems usually occur when sourcing different types of information or information from different sources especially when attempting to link environmental information with health effects and diseases:

- Differences in spatial and temporal resolution and coverage
- Differences in sampling/re-sampling and clustering methods especially when working with spatially as well as temporally measured and modelled data
- Differences in sensor/instrument accuracies
- Differences in parameter data characteristics for example the use of average versus maximum or minimum values or categorical versus numerical data or a daily integrated value versus a once off measurement made at a specific time of the day

4. IMPLICATIONS

These differences have a number of implications when linking or investigating the potential links between environmental (including in situ and satellite) and health (including case and laboratory) information.

Certain types of models have certain requirements for the type and format of input data needed. Exposure models generally require the concentration at specific intervals over a period of time when focusing on the effects of chemical and air pollutants. In general these pollutants have a cumulative health effect in that the prolonged exposure or number of exposures to critical threshold values lead to adverse health effects. For example the concentration value of a specific pollutant at hourly intervals for a 24 hour period and the number of exceedances of a critical threshold value within this period over a length of time is needed when using an exposure model for air pollution purposes. It will therefore be difficult or invalid to use an average or once-off daily satellite derived value as input for such an exposure model. It will however be possible to characterise the general air pollution conditions of the area studied over a period of time or at a specific time of the year and potentially link this with the number of people reporting respiratory problems. Exposure models focusing on the health effect of pathogens (i.e. viruses, bacteria and parasites) generally only require data on the concentration over time. The once–off exceedance of a critical pathogen concentration value can lead to an immediate health effect within a few hours, days or weeks. In some cases data on only the presence of the pathogen is sufficient.

The correlation or linking of two different time series datasets (e.g. an environmental and a health dataset) when using signal processing techniques is very limited when the temporal resolution differs between these datasets. This is especially true for instances when one dataset exhibits a high frequency signal that is not possible to detect in the second dataset due to resolution differences.

The differences in the spatial resolution and coverage between health and environmental data affect the quality and interpretation or even validity of any correlation results.

Differences in data quality influence the quality and interpretation of any analysis results.

Other complications associated with the correlation/fusion/linkage of health data with environmental data include:

- The restriction on individual patient information may necessitate a different presentation of the health data, e.g. clustering patient data on a larger spatial scale whereas the ground based measurements of air pollutants are very location specific.
- It is often not possible to establish a cause-effect relationship between a specific environmental variable(s) and a health condition or disease. Reasons typically include:
    - The combined effect(s) of more than one environmental variable, e.g. air pollutant concentration combined with ambient temperature and relative humidity, or the fact that people live in a rural area with no access to safe water and sanitation and during flood events they are exposed to waterborne pathogens
    - The relative contribution of the environmental parameter to the health effect or disease may change over time
    - The selection of the time period to be studied may have covered an above or below normal time period for a specific variable(s)

5. DATA FUSION METHODOLOGY

Data fusion for the purposes of this document refers to activities that determine the existence of a link, and its relative importance, between environmental (including climatic) factors and disease outbreaks and outbreak patterns over time and space. This is linked to the definition of data fusion presented by the Joint Directors of Laboratories (JDL), which formed the Data Fusion Subpanel (which later became known as the Data Fusion Group):

"A process dealing with the association, correlation, and combination of data and information from single and multiple sources to achieve:

- Refined position and identify estimates, and

- Complete and timely assessments of situations and their significance."

For the purposes of this document it involves the following categories and activities. These are not ordered in terms of importance or timing of execution.

5.1 Data Exploration

Data exploration includes the following activities:

- Identification of the most appropriate/suitable data set for a required parameter (based on criteria such as quality of the dataset estimated from a validation study or reported uncertainties; availability of the dataset for the time period of interest; applicability of the particular dataset to the area of interest for example using satellite data that employs an algorithm specifically developed for Case II waters (turbid, highly productive waters such as in river mouths and estuarine areas) versus a satellite product develop for Case I waters (clear water, e.g. in the ocean far away from the effects from land driven processes)
- Extraction of time series data for a selected area(s) based on criteria such as spatial and temporal coverage and/or resolution overlap with health data. Where overlap does not necessarily imply absolute locations in a Cartesian coordinate system. Areas that report cholera may be far apart in distance but the existence of a road network, water channel, or river linking the areas render them "near" as the disease can spread fast along these routes (via humans or water for example). Further, the selection of the location of the pixels on satellite images for the extraction of water-based parameters such as chlorophyll a concentration data are determined by the type of analysis envisaged. For example if the aim is to determine if phytoplankton plays a direct role in the outbreak of cholera in a coastal town or town next to a lake, pixels are most likely to be extracted for areas in the sea/lake close to the area reporting cholera. If the aim is to determine if phytoplankton only acts as a proxy (i.e. an indicator that captures environmental and climatic processes and factors that also affect the dynamics of the bacteria), pixels are likely to be selected within an estuary or river mouth or within a certain radius from the coastal/lake city that are reporting cholera or in the middle of the ocean outside the reach of the effects of land driven processes.

5.2 Data Preparation

Data preparation activities include:

- Reformatting of datasets, if necessary (i.e. data for the same parameter sensed at different time intervals by a sensor on a specific satellite may not cover the exact same area on the ground and the images need to be reformatted to adjust for edge effects or discontinuities, or different datasets have different projections or spatial and temporal resolutions). This needs to be done before these datasets can either be combined or analysed together. Remote sensing software tools such as the Modis Reprojection Tool, Erdas Imagine etc. can be used.
- Validation of remote sensing data, if necessary, using in situ data for the same environmental parameter, period and location.

5.3 Fusing Data

The fusing of environmental and health data does not produce a new data product but rather improves the current understanding of the disease and the links with the natural environment or produces predicted estimates of cholera case data using environmental data as input. This involves the analysis of a specific or combination of extracted environmental data in conjunction with extracted health data that can either be case data or microbiological laboratory results. Analysis techniques include:

- Time series analysis, e.g. ARIMA models and spectral analysis of individual data sets to determine trends and patterns in individual data sets as well as the existence of underlying signals (i.e. characterising data sets over time) within individual data sets
- Multivariate analysis, regression analysis and/or artificial neural networks to determine the most important environmental factors that can predict or are related to the presence of V. cholerae in water and sediment samples or the number of cases in a given area
- Correlation analysis, continuous wavelet transforms, crosswavelet analysis and fuzzy logic to determine relationships between two or more datasets over time.
- Spatio-temporal analysis, e.g. using techniques such as Self Organising Maps and Ripley's K index to cluster areas spatially based on the temporal characteristics of the health data. The main aim is to determine (if data are available and at a sufficient spatial and temporal resolution and coverage) the spread and rate of spread of outbreaks in an area.
- Statistical significance tests applied to all analysis.
- Linking laboratory results with remotely sensed or in situ data. The statistical analysis of the data produced when analysing for example water and sediment samples for microbiological, phytoplankton, and physiochemical properties are used to do the following:

    - Determine on the micro scale the most important factors driving the presence of the bacteria in the water samples and the sediment samples;
    - Identify and use possible available remote sensing or in situ data sets to track changes in identified factors;
- Uncertainty analysis. Health data are highly uncertain in terms of spatial accuracy. Case data usually do not reflect the area where the disease was contracted, only where it was reported. Case data can be underreported by as much as 90%. It is important to report information, if available, on the uncertainties associated with case data. The uncertainties associated with ex situ, in situ and remote sensing data also need to be reported. The error associated with final estimates when using the different data sets needs to be estimated. These uncertainties affect the analysis results but also the interpretation of results.

5.4 Model Development

Depending on the availability of data and data analysis results the following types of models can be developed:

- A statistical or artificial intelligence based prediction model to predict cases for a given area based on the results of the analysis activity. A specific or a combination of environmental factors are used as inputs
- Prediction model to predict the probability of an outbreak to occur or the risk in a given area based on a specific or a combination of environmental factors.
- Prediction model to predict the most likely start date of an outbreak based on a specific or a combination of environmental factors.
- Prediction model to predict the risk of spread to neighbouring areas over time.
- Simulation model to capture expert knowledge and the understanding gained from analysing the different datasets.

5.5 Validation and Testing

The availability of environmental and health data over space and time and real or near real time data will determine validation and testing activities. Typical validation and testing activities include:

- Validation of model results over time as new environmental, including climate as well as health data become available. In general historical data not used in analyses will be used for validation purposes during model development.
- Testing of model(s) developed for a specific location at different spatial locations.

5.6 Visualisation of outputs

The visualisation of outputs is important as it integrates complex data analysis results in an understandable way. Visualisation types can include:

- Dynamic maps that are regularly updated as new data become available showing areas that are at risk of an outbreak over time or static maps that show areas at risk due to historical environmental conditions over a period of time.
- Maps showing the spatial clustering of areas that are similar in terms of their response to environmental factors and cholera outbreaks
- Maps showing the spread and rate of spread of outbreaks in an area (provided cholera data are available on a fine spatial and temporal resolution)

6. CONCLUSIONS

Often, large amounts of environmental data are available and potentially accessible for processing by existent hardware and software tools. Just, the pure availability does not imply ‘easy’ and increased uptake by for example the health research community in general due to challenges mentioned. A systematic approach of which parts or most parts can be automated by using scientific workflow engines that integrate a number of processing steps and exempt researchers from the burden of manual execution and supervision at all steps of the scientific process. Other aspects, like sparseness of data in some areas will remain a problem for the time being, but might improve in the context of international initiatives such as GEOSS, the Global Earth Observation System of Systems. Still, it will take much more efforts to support the tracking of changes in health and natural environment in real time over large areas for data mining or adaptation of prediction models in most parts of the world.

ACKNOWLEDGEMENTS

This research was supported by the European Commission as part of the integrated project 'Earth Observation and Environmental modelling for the mitigation of Health risks' (EO2HEAVEN), contract no. 244100, and the Meraka Institute and Natural Resources and the Environment Operating Unit, Council for Scientific and Industrial Research, South Africa as part of provided Parliamentary Grants.

REFERENCES

De Magny, G.C. et al. 2008. Environmental signatures associated with cholera epidemics. *Proc. Nat. Acad. Sci. USA* 105 (46), pp 17676-17681.

OGC, 2010: OGC, 2010. OGC Fusion Standards Study Envineering Report. OGC 09-138

Rodó, X. et al. 2002. ENSO and Cholera: A Nonstationary Link Related to Climate Change? *Proc. Nat. Acad. Sci. USA* 99 (20), pp 12901-12906.

Woodborne, S., Pienaar, M., Van der Merwe, MR. 2008. Mitigating the future impacts of cholera. Conference proceedings: *Science real and relevant*: The 2nd CSIR Biennial Conference. Pretoria, South Africa, 17-18 November 2008. ISBN 978-0-7988-5573-0.

Zo, Y-G. et al. 2008. Covariability of Vibrio cholerae Microdiversity and Environmental Parameters. *Applied and Environmental Microbiology* 74 (9), pp 2915-2920.