Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

A novel approach for exposure assessment in air pollution epidemiological studies using neuro-fuzzy inference systems: Comparison of exposure estimates and exposure-health associations

Victoria Blanes-Vidal*, Manuella Lech Cantuaria, Esmaeil S. Nadimi

The Mærsk Mc-Kinney Møller Institute, University of Southern Denmark, Campusvej 55, DK-5230 Odense, Denmark

ARTICLE INFO

Keywords:
Air pollution
Health
Neural networks
Fuzzy logic
Exposure assessment

ABSTRACT

Many epidemiological studies have used proximity to sources as air pollution exposure assessment method. However, proximity measures are not generally good surrogates because of their complex non-linear relationship with exposures. Neuro-fuzzy inference systems (NFIS) can be used to map complex non-linear systems, but its usefulness in exposure assessment has not been extensively explored. We present a novel approach for exposure assessment using NFIS, where the inputs of the model were easily-obtainable proximity measures, and the output was residential exposure to an air pollutant. We applied it to a case-study on NH_3 pollution, and compared health effects and exposures estimated from NFIS, with those obtained from emission-dispersion models, and linear and non-linear regression proximity models, using 10-fold cross validation. The agreement between emission-dispersion and NFIS exposures was high (Root-mean-square error (RMSE) =0.275, correlation coefficient (r)=0.91) and resulted in similar health effect estimates. Linear models showed poor performance (RMSE=0.527, r=0.59), while non-linear regression models resulted in heterocedasticity, non-normality and clustered data. NFIS could be a useful tool for estimating individual air pollution exposures in epidemiological studies on large populations, when emission-dispersion data are not available. The tradeoff between simplicity and accuracy needs to be considered.

1. Introduction

Epidemiological studies often rely on exposure estimates to assess the association between exposure levels and health outcomes and to identify health risk factors. Health effect estimates and the conclusions arisen from epidemiological studies are influenced by these exposure data. It is therefore fundamental that epidemiological studies use reliable exposure assessment methods.

Different exposure assessment approaches can be used to estimate residential exposure to air pollutants. National routine monitoring networks provide long-term, nationally consistent air quality data, that can be used as a metric of exposure. However, these monitoring stations are often set for regulatory purposes and placed in locations that are not optimal for exposure assessment purposes (Özkaynak et al., 2013). More importantly, monitoring stations are expensive and hence very limited in number. As a consequence, exposure assessments based on monitoring stations lack spatial resolution and cannot provide accurate individual exposure estimates for pollutant concentrations that are spatially heterogeneous (Bell et al., 2011). Air pollution emission-dispersion modeling overcomes this difficulty by estimating pollutant concentrations at receptor sites based on emission inventories and meteorological data, using a complex set of governing equations describing the physical phenomena of emission, transport and fate of air pollutants. Air quality models can provide high spatial resolution, but they require complex physical models, precise information on source emissions, meteorology and topography, and they are costly and computationally demanding. Land-use regression models (LUR) can also be used to predict pollution at any location. LUR also has high input data requirements, since it utilizes the monitored levels of the pollutant of interest as the dependent variable and variables such as land cover, topography, and other geographic variables as the independent variables in a multivariate regression model (Ryan et al., 2007). Due to the high input data requirements of these methods, many studies have used proximity models, i.e. simple geographical features, such as distances to point sources or number of point sources within certain distance to the residence; as a proxy for exposure (Hodgson et al., 2007; Huang and Batterman, 2000; Vrijheid, 2000). Proximity measures are reasonably ascertainable and easy-to-use data, but it has been frequently discussed in the literature that using these measures as a proxy for exposure (i.e. proximity models) is not an

* Corresponding author.
E-mail address: vbv@mmmi.sdu.dk (V. Blanes-Vidal).

http://dx.doi.org/10.1016/j.envres.2016.12.028
Received 10 October 2016; Received in revised form 24 November 2016; Accepted 26 December 2016

accurate way of identifying exposed populations, since it can lead to significant exposure misclassification when compared with exposures estimated from atmospheric dispersion modeling and it may not provide reliable health effects estimates (Ashworth et al., 2013; Cantuaria et al., 2016; Cordioli et al., 2013; Hodgson et al., 2007; Kibble and Harrison, 2005). Due to these risks, previous studies have come to the conclusion that in order to obtain accurate and reliable estimates of residential exposure, it is imperative to have precise data on emission, meteorology and topography data and to use emission-dispersion modeling or LUR (Hodgson et al., 2007; Özkaynak et al., 2013).

Given the shortcomings of emission-dispersion modeling (i.e. requires extensive input data) and methods based on proximity to point sources (i.e. lack of accuracy), it would be interesting to develop alternative methods able to estimate air pollution exposures at receptor sites for large populations with high accuracy and spatial resolution (similar to those obtained from emission-dispersion modeling approaches), but do not have high input data requirements. The final goal of such exposure estimation method is prediction, which is obtained by mapping a set of variables (i.e. easy-measurable input variables, such as proximity measures) in input space to a response variable (i.e. air pollution exposures estimates at receptor sites) in the output space through a model. Mathematical approaches for developing these models must consider that the cause-effect relationships of these parameters leading from air pollutant emission to individual exposure, are complex, uncertain, and non-linear in nature (Zou et al., 2016).

Fuzzy logic (i.e. the logic of fuzzy sets) is suitable for uncertain or approximate reasoning, especially for complex non-linear systems with a mathematical model that is difficult to derive. A fuzzy inference system (FIS) can be defined as the nonlinear mapping of an input data set to a scalar output data, using fuzzy logic (Mendel, 1995). Fuzzy logic is based on fuzzy sets and fuzzy rules. Fuzzy set theory can be conceptualized as a generalization of classical Aristotelic logic. In classical logic, the membership of elements in a set is assessed in binary terms (i.e. objects either belong to a set or do not belong to a set). Therefore classical logic only permits propositions having a value of truth or falsity. Fuzzy sets, unlike classical sets, are defined by membership functions. A membership function assigns to each element in the set under consideration (the universal space) a degree of membership to the set, which is a value in the interval [0,1]. Therefore, an element can simultaneously belong to several subsets, at least to a certain degree of membership. In fuzzy logic, the description of how the FIS makes a decision regarding the output, based on the inputs, is expressed by a collection of linguistic IF-THEN statements called fuzzy rules.

Fuzzy inference systems have as a limitation that they do not have the ability to learn from the data, and therefore, the successful performance relies heavily on human knowledge derived from domain experts who, based on their experience, define the shape of the membership functions and define the fuzzy rules. Introducing artificial neural network (ANN) data driven optimization techniques improves the potentially inaccurate or incomplete information provided by the experts. In a neuro-fuzzy inference system (NFIS), the system is trained by means of a data-driven learning method derived from neural network theory. Fuzzy logic and neuro-fuzzy inference systems are well-established concepts in mathematics and engineering but its usefulness in medicine, epidemiology, health risk assessment and exposure assessment has not been extensively explored (Giubilato et al., 2014; Godil et al., 2011; Massad et al., 2003; Milla dos Santos et al., 2014; Vineis, 2008).

Neuro-fuzzy inference systems have been used in air pollution studies to create air quality indexes using specific air pollutant concentrations (e.g. $PM_{2.5}$, PM_{10}, O_3, CO, NO_2 and SO_2) as input variables (Olvera-García et al., 2016). Besides, ANN and NFIS have been used to estimate air pollution concentrations in two main applications. First, for air pollution time series forecasting, using as input variables: previous concentrations of the air pollutant being investigated, concentrations of other relevant air pollutants and meteorological variables, at the location of interest (Dursun et al., 2015; Morabito and Versaci, 2003; Zahedi et al., 2014). In relation to air pollution forecasting, Ausati and Amanollahi (Ausati and Amanollahi, 2016) recently used NFIS to predict $PM_{2.5}$ based on SO_2, PM_{10}, O_3, $PM_{2.5}$ on the previous day, average maximum temperature and wind speed; and Mishra and Goyal (Mishra and Goyal, 2016) proposed an artificial intelligence based neuro-fuzzy model for NO_2 forecasting using as inputs: temperature, pressure, relative humidity, wind speed, wind direction index, visibility and previous day's NO_2 concentrations (estimated from an emission-dispersion model). Second, ANN and NFIS have also been used for building spatially dense air pollution maps based on known air pollutant concentrations at specific locations. Regarding spatial estimations, Shahraiyni et al. (2015) used artificial neural networks to estimate the hourly PM_{10} concentration at monitoring stations that have been shut-down in Germany, based on PM_{10} measurements of still–operating monitoring stations; while Wahid et al. (2013) estimated the spatial distribution of daily O_3 concentrations at the state of New South Wales (Australia), by building a neural network model that approximates the nonlinear relationship between NO_x emission, ambient temperature, location coordinates and topography, considered as the inputs, and the 8-h maximum average of O_3 concentration as the output. To sum up, previous NFIS models are able to successfully estimate air pollution concentrations, under condition of possessing a proper database of data, typically a series of air pollution and meteorological variables. However, to our knowledge, no previous study has aimed at using NFIS to estimate air pollutant concentration from simple proximity measures (e.g. distance to sources).

In this study we investigated neuro-fuzzy inference systems as a methodology to estimate residential exposures to air pollutants, using simple proximity measures (i.e. number of point pollution sources within certain distances to the residence and in the region), as sole model inputs. We described this methodology, applied it to a case-study and evaluated the accuracy of the neuro-fuzzy method for exposure prediction. We further compared air pollution exposures predicted by NFIS with those obtained from linear and non-linear regression methods that use the same proximity measures as input variables, as in NFIS. Finally, we evaluated the effect that using these different exposure assessment methods (i.e. emission-dispersion models, linear and non-linear regression proximity models and NFIS proximity models) may have on the health effect estimates.

2. Materials and methods

2.1. Fuzzy inference systems

The NFIS architecture can be described as a structure of six different layers (Fig. 1), which represent the process from the model inputs to the predicted output. In our study we used a Sugeno-type inference system, where the functioning of each layer is as follows (Abraham et al., 2002):

Layer 1 is the input layer. No computation is performed here, each node in this layer only transmits the inputs to the next layer.

Layer 2 is the fuzzification layer. Fuzzification is a procedure through which the input variables are turned into the degrees of membership to given fuzzy sets or classes, as determined by membership functions. The fuzzification layer contains information on the membership functions of input variables, which can be any appropriate parameterized function introduced in here. Typical membership functions are triangular, trapezoidal, simple Gaussian curve, two-sided composite of two different Gaussian curves and the generalized bell membership function (Supplementary information). The outputs of this layer are represented as $\mu_{A_{j,l}}(x_j)$ which is the degree of membership