# Air pollution and population morbidity forecasting with artificial neural networks

**A Yu Gornov[1], T S Zarodnyuk[1] and N V Efimova[2]**

[1] Institute for System Dynamics and Control Theory, Siberian Branch of Russian Academy of Sciences, Lermontov str. 134, Post Box 64033 Irkutsk, Russia
[2] East-Siberian Institute of Medical and Ecological Research, 12a District, 3, 665827 Angarsk, Russia

**Abstract**. Incidence prediction models for urban population have not yielded consistent or highly accurate results. The complex nature of the interrelationship between "environmental factors and incidence" has many nonlinear associations with outcomes. We explore artificial neural networks (ANNs) to predict the complex interactions between the risk factors of incidence among the urban population. ANN modeling using a standard feed-forward, back-propagation neural network with three layers (i.e., an input layer, a hidden layer, and an output layer) is used to predict the incidences of diseases of children and adults. A receiver-operating characteristic (ROC) analysis is used to assess the model accuracy. We develop a mathematical model taking into account factors of natural, anthropogenic, and social environments. The model effectiveness is proved by computing experiments for the Bratsk industrial centre (Irkutsk region, Russia). Optimal air pollution levels are offered to achieve a background morbidity level among different age groups of the population. The prediction of incidence is most accurate when using the ANN model with several univariate influences on the outcome. An incorporation of some computerized learning systems might improve decision making and outcome prediction.

**Keywords:** incidence, prediction, outcome assessment, computer simulation, environmental factors, artificial neural networks

## 1. Introduction

The development and analysis of mathematical models for describing the dynamics of interaction between the system elements are used to assess and forecast the state of a medical-ecological system. Mathematical models allow us to coordinate different information, which is necessary for any complex interdisciplinary research [1, 2]. In addition, in terms of evidence-based medicine the application of mathematical, statistical, and epidemiological approaches is absolutely necessary [3, 4]. In recent years the modeling of complex processes and systems have used neural network techniques, which are based on the idea of building a computing device of a large number of simple elements operating in parallel (the so-called formal neurons). These artificial neurons function independently of each other and are connected by channels of communication [5, 6]. Each individual neuron is simulated of a simple logistic function, but high complexity of the model and the flexibility of its operation are determined by the connections of the structure and a multi-level hierarchy of the network. The neural network taking, as input, a signal can put out a definite answer, which depends on the weight coefficients of all artificial neurons in the process of self-learning network [7].

The artificial neural networks have become very popular in various scientific investigations. The mathematical results obtained by Kolmogorov, Arnold, Stone, and Gorban allowed one to develop a large number of software and computing techniques [2, 5, 8]. Compared to the traditional regression

approach, the ANN is capable of modelling complex nonlinear relationships. The ANN also has an excellent fault tolerance and is fast and highly scalable with parallel processing. However, in the authors' opinion, for every specific problem it is necessary to solve a number of important questions: choose a neural network structure, a number of layers and neurons, the type of activation functions of neurons, which almost no universal methods. An unsatisfactory solution of the above sub-tasks can lead to the construction of bad approximations and completely wrong forecast results.

The purpose of this investigation is to estimate the use of artificial neural networks for studying the characteristics of the morbidity origin for a large industrial centre.

## 2. Materials and methods

The following tasks have been solved in this article:

- development of a mathematical model that takes into account the factors of the natural and social environments;
- determination of the background levels of morbidity based on a neural model with the most significant factors of the environment;
- elaboration of the optimal levels of air pollution, which gives the background levels of morbidity of various age groups;
- verification of the model by comparing the results of computing experiments obtained by different techniques.

The studies were conducted in the Bratsk industrial centre located in the north of Irkutsk region (Russia). In 1994 the results of a state expert review of the environmental conditions and population health in the city were recognized as critical. The introduction of the federal program of urgent measures to improve the environmental, sanitary, and epidemiological situation and the health of the Bratsk industrial centre in 1995–2005 can be considered as a major experiment realized by using modern mathematical models and software.

It is necessary to solve one of the difficult problems for a qualitative prediction of the population health in the region. This problem consists in finding the environment-related component in morbidity and constructing a mathematical model describing the population health.

The behaviour of the model elements is estimated in the interconnection between the individual elements and the "factors – response" cause-effect relationship. We took into account the dependence of the air pollution level from the weather conditions by using the results of numerous studies presented in [2, 4, 9]. A mathematical formalization for the dependence of the population morbidity $Z_i$ on the most influential factors is as follows:

$$Z_i = a_1 \cdot T_i + a_2 \cdot W_i + a_3 \cdot V_i + a_4 \cdot \frac{LnHI_i}{LnV_i} + a_5 \cdot LnC_i , \quad (1)$$

where $T_i$ is the average annual air temperature (°C); $W_i$ is the outpatient help provision (the number of physicians per 1 000 population); $V_i$ is the average annual wind speed (m / s); $HI_i$ is the integrated index of air pollution (standard units); $C_i$ is the level of social conditions in the city (expert review, points); $a_k$ is the coefficient, $k = \overline{1,5}$. We considered the time interval from 1990 to 2005. This period was characterized by a pronounced change in the quality of the environment and public health.

We took into account the complex nature of formation of the morbidity rates, which is defined as a set of long-term and short-term effects (chronic and acute ones, accordingly) of various factors. The annual average indicator of the total air pollution from complex substances, *HI*, calculated in accordance with [10] is used as a measure of the long-term chemical contamination value. The assessment of air pollution is based on the number of pollutants released into the air, and their physical, chemical, and toxicological properties, and the possibility of potentiation and addition of the biological influence. The indicators of overall morbidity were studied with a sample according to the data from medical institutions. The morbidity rates are based on the average value (*M*), its error (*m*),

and 95% confidence intervals (CI). The information about the social conditions submitted for evaluation to experts included the quality of housing (the availability of centralized water supply and sewerage, electric stoves, the average living space per capita), the social infrastructure (availability of kindergartens, schools, cultural and sports institutions), and some characteristics of living obtained by different authors in the survey of the population.

An approximation of the considered function of five variables by an artificial neural network is one way of constructing a predictive model. A general training sample for the population of ten cities of Irkutsk Region according to the statistical reports of the Committee on Statistics, health care institutions, Centres of Hygiene and Epidemiology, the Department for Hydrometeorology and Environmental Monitoring was created to construct the mathematical model. The medical, social, and environmental data (1990–2005) used it in a training/cross-validation group. The data of the internal validation group included indicators from 2006 to 2016. The training/cross-validation group was used to train the network. Several methods are used for training the network (internal adjustment weights network for the best approximation of the considered dependence), there are the least squares method, the algorithm of random multistart, the simulated annealing and conjugate gradient methods of Fletcher-Reeves and Polak–Ribiere [5]. The computational technology was developed by the authors to construct a backpropagation network. This software is a neuro-emulator and is based on the above approaches.

## 3. Results

The models describing the dynamics of morbidity of children and adults for the Bratsk industrial centre are obtained as a result of computational experiments and the performed procedure of parametric identification of the mathematical model (1). To solve these problems we used a personal computer with a processor Intel Core i5-2500K and 16 Gb RAM.

The best results were obtained with a two-layer neural network with 5 and 10 neurons in respective layers (Table 1). The computations for the network with 20 neurons in layers provided the least accurate solution. The estimated morbidity is less than the annual average actual rate for children 4.9–16.1 times, and for adults 4.8–13.5 times. It should also be noted that the computing time for different neural network structures was notable for slightly and was 386–420 sec for analysing the children morbidity data, and 402–468 sec for computing the morbidity of adults.

**Table 1.** Computing results of morbidity for Bratsk industrial centre obtained by artificial neural networks.

| Number of neural network layers | number of neurons in each layer | Children | | Adults | |
|---|---|---|---|---|---|
| | | Morbidity, ‰ | Computing time, sec | Morbidity, ‰ | Computing time, sec |
| 1 | 5 | 1762.2 | 420 | 1039.2 | 420 |
| 1 | 10 | 1638.4 | 420 | 1131.1 | 420 |
| 1 | 20 | 1611.7 | 420 | 1080.2 | 420 |
| 2 | 5/5 | 1576.7 | 401 | 833.0 | 468 |
| 2 | 10/10 | 1614.7 | 420 | 1067.2 | 420 |
| 2 | 20/20[a] | 368.3 | 420 | 277.3 | 423 |
| 3 | 5/5/5 | 1618.7 | 386 | 869.1 | 402 |
| 3 | 10/10/10 | 1154.5 | 420 | 991.5 | 420 |
| 3 | 20/20/20[a] | 108.4 | 430 | 97.9 | 443 |

[a] Note: The estimated value is unsatisfactorily approximated data.

The only way to assess the reliability of the calculations with different neural networks is neuro-consulting. To use this technique, several networks are trained for a solution to the same problem. We chose the average of the obtained results as a response. The accuracy of the final result is estimated by the deviation from the mean response. To predict the morbidity rate of the Bratsk city population, we used neuro-consulting. It consists of 15 neural networks trained independently. The result of prediction is defined as the average of the forecast results of the consulting networks. The training of the neural networks consulting is to predict the morbidity for certain values of the parameters with the application of the statistical information accumulated over 10 years. The developed neural model allows us to predict the morbidity of children, adolescents, and adults for certain values of the initial parameters: $T_i = -3.5$ °C; $W_i = 2.8$ physicians per 1000 people; $V_i = 2.6$ m/s, $HI_i$ = 15.0 standard units; $C_i = 8$ points. The results of the predictive calculations of the neuro-consulting are presented in Table 1.

We verified the results with a test period corresponding to the problem parameters. 1999–2001 proved to be the nearest to the initial data. In this time period there are relatively low levels (for the Bratsk city) of air pollution (*HI* = 12–16) with a stable climate and social factors. The morbidity is rather an inertial measure, and often it does not react to changing conditions during the investigated calendar year, and therefore we took into account the average morbidity rate for three years for each age group. The averaged computing data consists of 1554 children (95% of CI is from 1 287 to 1 827) per 1 000 people (‰), and 1020 ‰ adults (95% of CI from 787 to 1 253). We found that the actual morbidity for children exceeds the upper limits of the computing data by 18%, and for adults it is about 17%, the mean differences are not statistically significant ( $p > 0.05$ ).

Mathematical and information techniques for controlling complex systems have been traditionally used. They allow, by applying models, to consider the implications of introduction of a controlled decision [1]. In connection with the above, we were interested in the problem of determining the optimal variables in order to achieve the target morbidity rates. The target indicators of any social program can serve as morbidity levels in the region, for Irkutsk region it is 1300 ‰ for children and 1200 ‰ for adults. We took into account the probabilistic limits of variability of the predictors in the calculations. These limits are based on the empirical observations of the 30-year period. The boundary values of the natural conditions were as follows: the temperature changed between 0 and –3.5 °C, and the wind speed between 1.5 and 2.8 m/s. The change limits of controllable facts were chosen to achieve good quality of life in the following way: the provision of medical doctors is below the average value for the Russian Federation and above it for Irkutsk region (2.0–3.0 per 1 000 population), the social conditions require no additional investments and maintaining the lifestyle of the majority of the population (it is estimated to be 6 points), or the social conditions improve with healthy lifestyle (10 points). The lower limit of the integral index of air pollution is selected within the tolerance level (*HI* is changed from 2 to 4). We also considered the upper limit *HI* – 15 which represented high pollution.

The forecast results are presented in Table 2. It was found that morbidity largely depends on the environmental pollution and can be achieved even in more severe weather conditions. To reduce children morbidity, it is very important to improve health care and reduce air pollution. It should be noted that reducing the air contamination for a city with more than 30 industrial enterprises (including an aluminum plant, facilities of thermal energy and pulp paper and chemical industry) to that level is really possible when implementing a number of environmental measures. These computing results were the basis for the proposals considered in the development of the municipal program for the conservation of the environment and the city population in 2008–2012.

**Table 2.** Computing results to achieve optimal levels of morbidity for the population of Bratsk industrial centre.

| Parameter | Optimal level | |
|---|---|---|
| | Children | Adults |
| Average annual air temperature, °C | -3.5 | -3.5 |
| Physicians provision, ‰ | 3.5 | 2.8 |
| Annual average wind speed, m/s | 2.6 | 2.6 |
| Level of social conditions, points | 7.5 | 4 |
| Air pollution index, standard units | 9 | 7 |
| Morbidity ($M \pm m$),‰ | 1312 ± 40 | 998 ± 36 |

## 4. Discussion

Many types of neural networks have been designed already and new ones are invented every week, but all can be described by the transfer functions of their neurons, by the learning rule, and by the connection formula. In terms of the model specification, artificial neural networks require no knowledge of the data source but, since they often contain many weights that must be estimated, they require large training sets.

This study demonstrated that the use of ANNs to enhance the outcome predictions regarding medical-environmental dependents had any limitation. The computation by neural networks with 20 neurons in each layer gives the least accurate result. This is probably due to the fact that the time interval of the discussed data is quite short [2]. Thus, the use of the present network structure in epidemiological studies based on annual reports for periods of time less than 20 years is not acceptable. To assess the influence of factors on certain subpopulations, the presented models can be extended. It is necessary to consider the changes in the production conditions in study of the formation of adult population morbidity. This chemical factor has a complex impact on the greatest of the adult population [3, 11]. We are going to perform a series of computing experiments to explore the features of the constructed model, to adapt it to the specific conditions, and to create short- and medium-term situational forecasts. Note that the ANN models suggested a difference in the weighted importance of the included factors that changed based on how the variables were included and combined in the model. Based on the differences in the model inputs and the resulting outputs, the outcome predictions of incidences were complex and likely to be influenced by several factors that might have unknown relationships with each other (which cannot be easily demonstrated using standard near-linear regression models). The cumulative presence of unknown and small relationships might largely explain the difficulty, limited success, and moderate accuracy associated with generating predictive models consisting of a few (e.g., 3–5) explanatory variables, given that the actual outcome depends on a much greater variance and (at least partially unknown) interdependence of the factors. In reality, many subtle factors may indeed point to a potentially dire situation that may easily be missed or unrecognized by the inexperienced clinician. Situations like these are where computerized pattern recognition and prediction algorithms can become useful [12]. Although their full potential has yet to be reached [13], developments in technology are rapidly moving toward models that might become available for everyday use.

ANNs gather their knowledge by detecting patterns and relationships in data and learn (or are trained) through experience, but not from programming. An ANN is formed from hundreds of single units, artificial neurons or processing elements, connected with coefficients (weights), which constitute the neural structure and are organized in layers. The weights are the adjustable parameters and, in this sense, a neural network is a parameterized system.

The application of neural network techniques in the study of complex biomedical systems allows one to solve the problems of classification and minimizing the number of required forecast parameters. These problems constantly arise in the differential diagnosis in clinical medicine and in clustering areas of medical and environmental situations for socio-hygienic monitoring.

There are several limitations in this study. First, the city of Bratsk was not large enough to solve our problems. We had to use data on other cities in Irkutsk region to increase sample representativeness. However, the generalizability and reliability of the findings are a concern. Second, rapid and not-well-planned urbanization is associated with a high level of ambient air pollution, mainly caused by industrial sources and vehicular exhausts. There is sufficient evidence of the adverse effects related to short-term exposure, while fewer studies have addressed the longer-term health effects [14]. Increased pollution exposures have been associated with increased mortality, morbidity, hospital admissions/emergency-room visits, mainly due to exacerbations of chronic diseases or respiratory tract infections [15–18]. These effects may also be modulated by the ambient temperature, and many studies show that the elderly are most vulnerable to heat waves [19, 20]. The state data about the population health were a precise data source for determining the status and trends in the occurrence of different diseases. Third, since socioeconomic data were estimated by experts, they had limitations for different studies. Fourth, because the period of data collection spanned over a long time, there may be some bias related to disease diagnosis caused by changes in the policy, as well as changes in the availability of new diagnostic or analytical methods.

**5. Conclusions**

We showed that the above mathematical model of "disease – environmental factors", which was constructed with the use of neural networks in a time interval of 16 years, allows fairly accurate predictions and management decisions to minimize risks to public health. The potential applications of the ANN methodology in interdisciplinary research range from interpretation of analytical data to the creation of a program of monitoring and control. In addition, ANNs can combine and incorporate both literature-based and experimental data to solve the problems. The experience obtained in the perturbation action setting to achieve a population health indicator can be useful in programs aimed at improving the natural and social environment and the preservation of public health.

**Acknowledgments**

This work was partly supported by the Russian Foundation for Basic Research, Grant N 17-07-00627.

## 6. References

[1] Gurman V I and Ryumina E V 2001 *Modeling socio-ecological-economic system of the region* (Moscow: Nauka) p 175

[2] Nuterman R, Starchenko A, and Baklanov A 2011 Numerical model of urban aerodynamics and pollution dispersion *International Journal of Environment and Pollution*, **44**(1-4) 385–93

[3] Budtz-Jørgensen E, Debes F, Weihe P, and Grandjean P 2010 Structural equation models for meta-analysis in environmental risk assessment *Environmetrics* **21**(5) 510–27

[4] Elangasinghe M A, Singhal N, Dirks K N, and Salmond J A 2014 Development of an ANN-based air pollution forecasting system with explicit knowledge through sensitivity analysis *Atmospheric Pollution Research* **5** (4) 696–708

[5] Gorban A N and Rossiev D A 1996 *Neural network on a personal computer* (Novosibirsk: Nauka) p 278

[6] Zou J, Han Y, and So S 2008 Overview of artificial neural networks *Methods Mol. Biol.* **458** 15–23

[7] Suykens J A K, Vandewalle J P L, and De Moor B L R 1996 *Artificial neural networks for modelling and control of non-linear systems* (New York: Springer) p 316

[8] Cheng S Y, Li L, Chen D S, and Li J B 2012 A neural network based ensemble approach for improving the accuracy of meteorological fields used for regional air quality modelling *Journal of Environmental Management* **112** 404–14
[9] Singh K P, Gupta S, Kumar A, and Shukla S P 2012 Linear and nonlinear modelling approaches for urban air quality prediction *Science of the Total Environment* **426** 244–55
[10] Risk assessment. Guidance for Superfund 1989 *Human Health Evaluation Manual* 1 (Washington: Environmental Protection Agency) p 291
[11] Blair A, Stewart P, Lubin J H, and Forastiere F 2007 Methodological issues regarding confounding and exposure misclassification in epidemiological studies of occupational exposures *Am J Ind Med* **50** 199–207
[12] Waljee A K, Higgins P D R, and Singal A G 2014 A primer on predictive models *Clin Trans Gastroenterol* **5** e45
[13] Jones N 2014 Computer science: the learning machines *Nature **505*** 146–8
[14] Furukawa M F, Poon E 2011 Meaningful use of health information technology: evidence suggests benefits and challenges lie ahead *Am J Manag Care* **17(**12 Spec No.**):**SP76a-SP
[15] Kamińska J A 2018 The use of random forests in modelling short-term air pollution effects based on traffic and meteorological conditions: A case study in Wroclaw *Journal of Environmental Management* **217** 164–74
[16] Paciorek C J and Liu Y 2012 Assessment and statistical modelling of the relationship between remotely sensed aerosol optical depth and PM2.5 in the eastern United States *Res Rep Health Eff Inst* **167** 5-83
[17] Girguis M S, Strickland M J, Hu X, Liu Y, Chang H H, Belanoff C, Bartell S M and Vieira V M 2017 Chronic $PM_{2.5}$ exposure and risk of infant bronchiolitis and otitis media clinical encounters *Int J Hyg Environ Health* **220**(6) 1055–63
[18] MacIntyre E A *et al* 2014 Air pollution and respiratory infections during early childhood: an analysis of 10 European birth cohorts within the ESCAPE Project *Environ Health Perspect* **122**(1) 107–13
[19] Simoni M, Baldacci S, Maio S, Cerrai S, Sarno G, and Viegi G 2015 Adverse effects of outdoor pollution in the elderly *J Thorac Dis* **7**(1) 34–45
[20] Revich B A and Shaposhnikov D A 2016 Cold waves in southern cities of European Russia and premature mortality *Studies on Russian Economic Development* **27** 210–15