Pisco Med Publishing

Original Research Article

Analysis on Risk Factors and Spatial Variation of Pollution Sources in Urban Areas

Zheng Xu[1], Qi Zhu[2]

1. College of Architecture and Urban Planning, Tongji University, Shanghai, China, 200092

2. Chinese Academy of Sciences, Key Laboratory of Forest Ecology and Management, Beijing, China, 100000

Abstract: In this paper, two mathematical models are used to discuss the impact of urban pollutants on human living environment, and it is concluded that pollutants are the main cause of environmental pollution and have no direct relationship with pollution sources. Through the data of human health indicators and fitting the data sets of the two models, we want to investigate whether there are any variables related to the increase of skin cancer risk and whether the increase of cancer risk is related to the distance from coal-fired power stations in the Slovakian town of Novaky. In this work, we comparing two models, it indicates that environmental and urinary arsenic have significant impact on the cancer risk, but the spatial location cannot be consider as a direct factor on non-melanoma skin cancer.. By comparing both models, it indicates that environmental and urinary arsenic have significant impact on the cancer risk, but the spatial location can not be consider as a direct factor on non-melanoma skin cancer. The project case, the specific planning and engineering design, should focus on strengthening the control of the pollution source, rather than merely considering spatial layout of pollutants.

Keywords: Analysis; Risk factors; Spatial Variation; Pollution Sources; Urban Areas.

1. Introduction

1.1 Background

Non-melanoma skin cancer (NMSCs) constitute more than one-third of all cancers in the U.S. NMSCs are the most common malignancies occurring in the white populations each year. Most cases are caused by over-exposure to UV rays from the sun or sun beds. NMSCs mainly refers to base cell carcinoma, followed by squamous cell carcinoma (L,D.T.& V,M., 2002).

A case-control study was set up from 1996 to 1999 in order to identify the associations between arsenic exposure and non-melanoma skin cancer incidence in the vicinity of a coal-burning power station in Slovakian town of Novaky. There were 257 cases of non-melanoma skin cancer were identified within the region of study, defined as a circle of radius 25 km, and 211 healthy controls living in the region were selected for control (Bodrud-doza,M.Islam,A.T.Ahmed,F. et al., 2016).

1.2 Dataset

The dataset that collected from this case-control records the following information: both access and study ID of the test population (ID and idn); case or control status (caco); sex and age of the test person (sex and age); distance category in the original analysis (distance) (Zhongmin,J., Siyue,L.& Li,W., 2018); smoking status (smoke2); environmental arsenic which based on power station emissions (ares1); the x y coordinates (x and y); name and ID of the study towns (sat place and ressta); soil arsenic at case r control address and total urinary arsenic (soilas and sumas)(Wen-cong,L., Ning-lu,C.& Wen-xin,Q., 2017).

N.B: all above factors are given in short names in the dataset as shown in the brackets.

1.3 Aim of this project

The dataset has include a measure of arsenic in the soil at the residential address of each individual, and a measure of urinary arsenic for each individual, this project will be interested in the identifications of any of the recorded variables are associated with increased risk of non-melanoma skin cancer and the investigation of whether increased cancer risk is associated with distance from the power station.

2. Methods

This section will introduce brief ideas of how to fit regression model and isotropic Gaussian model to the dataset , so that the association between cancer risk and the two arsenic measures can be investigated, and find possible confounders as appropriate (Chang,X.Waagepetersen,R.Yu,H. et al., 2015).

2.1 Fit generalised linear model to the data

Without considering the spatial locations, we can fit a generalised linear model to the dataset and deduce the associated risk factors to the non-melanoma skin cancer (Zhu,G., Ge,Y.& Wang,H., 2013). As we can see that each test subject either has got non-melanoma skin cancer or not, it is naturally consider each test subject as an iid Bernoulli trial, so the generalised linear model to fit this

doi: 10.18282/l-e.v10i3.2406

dataset can be considered under binomial family, and the expression is:

$$log\frac{p}{1-p} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots \qquad (1)$$

Where β_i , s are coefficients and (x_{11}, x_{12} , x_{13},...) are the covariates, in this case, these covariances refers to the risk factors that associated to non-melanoma skin cancer.

From the dataset we notice that there are 7 covariances that we are going to consider (sex, age, distance, smoke2, ares1, soilas and sumas), it is possible that some of the factors do not link to the increase of the risk to get non-melanoma skin cancer, so that we will need to find out which factors will really contribute to the cancer risk and the association between cancer risk and the two arsenic measures (soils and sumas). This model will help to identify the important risk factors of non-melanoma skin cancer, but we also like to investigate whether increased cancer risk is associated with distance from the power station. So the following model will be introduced (Ahmed, N.Bodrud-doza,M.Islam, A.R.M.T. et al., 2019).

2.2 Fit the isotropic Gaussian model to the dataset

In some researches, we have noticed there are some findings on the impact of location and the risk of getting a particular disease. Within these situations, the risk of getting a particular disease will be affected by the distance between the location of test subjects and a point source. Generally speaking, the risk of get a disease might vary with the distance to a point source, i.e. near or far. In this project, we like to find out that whether the coal-burning power station in Novaky is a point source that affect the risk of getting non-melanoma skin cancer of the people live around it. Unlike the above model, this model will have to take count of the spatial locations. As we know the test region is defined as a circle of radius 25km centred the coal-burning power station, with location (18.5334, 48.6993), we like to transfer the spatial variables to planar co-ordinates. Use the coordinates of the power station as the origin, the transformation equation is

$(x_{i,new},y_{i,new})= \{(x_i\text{-}18.5334)*73.611,(y_i\text{-}48.6993)*111.204\}$

This equation is obtained by the approximation method of National Geospatial Intelligence Agency [2] .

So $s = \sqrt{x_{i,new}^2 + y_{i,new}^2}$ measures the distance between the individual's location and the coal-burning power station.

Now we consider ρ as a function represents spatial variation in risk, the spatial variation in risk of a distance d from a point source is then $\rho\,(d) = 1 + \beta\exp\left\{-\left(\frac{d}{\delta}\right)^2\right\}$

Where β is elevation in risk at source and δ is the rate of decay of risk with distance from source, as in the isotropic model.

3. Pre-step on data analysis

In this dataset, there are lots of missing values, table 1 and 2 give out the total number of missing values in rows and columns. From Table 1 we can conclude only 183 rows with no missing values, and table 2 show the columns of soilas and sumas also has lots of missing values. As these two columns contains important information that we need to test on, so with lots of missing values could lead to bias in the analysis.

Table 1: Counts of the NA's in each raw of the dataset

No. of missing values	0	1	2	6
Count	183	262	15	8

Table 2: Counts of NA's in each row

Variable	caco	sex	age	distance	smoke2	ares1	soilas	sumas	x	y
No.of NAs	0	0	0	0	1	0	262	29	0	0

Figure 1, 2 & 3 give us a look of relations between risk factors , case, control and locations. From Figure 1, we can see that clearly patterns from both plots, which means that most test subjects have similar soil and urinary arsenic measurements even though varying with distance from the power station.

Figure 2 shows us cases and controls within the study region, we can see that the centre of both circles is the power station, and from the intensity of the points we can tell that the distance from the most recorded cases is not the shortest to the power station and similarly for control plot.

Figure 3 shows us the relative risk in both types of plots, and we can tell that around the power station, it is true that the relative risk to get non-melanoma skin cancer is higher, but the northeast part of our test region also have high relative risk to get skin cancer even though there are comparable far away from the power station. Both figure 2 and 3 might shows a result that the cancer risk might not tightly associated with the distance from the power station, and hypothesis will be tested later on by using Diggle-Rowlingson model.

4. Result

4.1 Fit dataset to generalised linear regression models

As we are aiming to investigate not only the association between cancer risk and the two arsenic measures but also adjusting for possible confounders, we will test all 7 variables together for the first time.

From the R output above, with AIC=254.62, we can tell that most of the covariates make this regression model become insufficient. If we are testing under 10% significant level, only ares1 with p-value 0.0751 and sumas with p-value 0.0970 which are less than 0.1, so other factors might not be considered as significant factors that associated to cancer risk. All these should suggest an improvement of the regression model with selected variable.

Now we conclude a new regression model with 3 variables: ares1, soilas and sumas.Still from the R output above, we can clearly see that both ares1 and sumas are risk factors that we are looking at, but still soilas does not shown to be an associated factor to cancer risk, as it has high p-value=0.6221, so it could be a better idea to remove soilas and fit our final model with ares1 and sumas. Our final model becomes, from Equation (1): $log\frac{p}{1-p} = -1.36674 + 0.65944 \cdot ares1 + 0.04662 \cdot sumas$

This better model is with p-value=0.073, it turns out that urinary arsenic is a factor that associated to cancer risk and the confounder is the environmental arsenic which is based on power station emissions. Our initial guess of one risk factor, soil arsenic is not a major factor that associated to the risk of getting non-melanoma skin cancer.

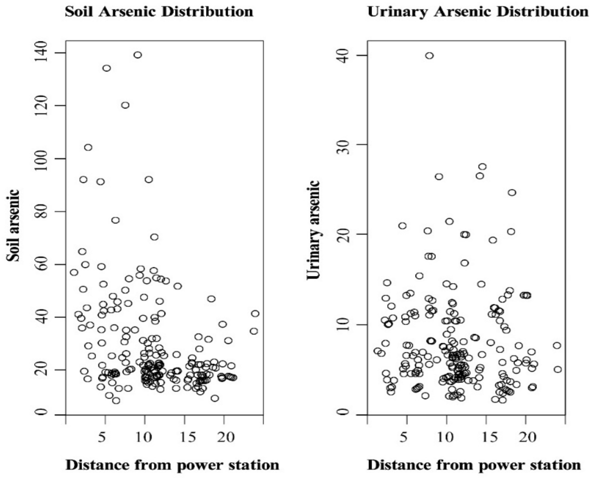

Figure 1: Plots of the two arsenic measurements against distance to the power station

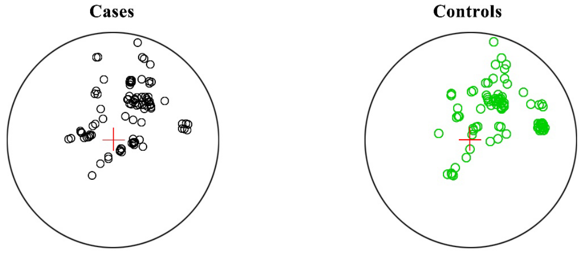

Figure 2: Location of cases and controls, centred with the power station

4.2 Fit dataset to Diggle-Rowlingson model

We have discover that soil arsenic does not count to be associated to the cancer risk, and from table 1 we see that it is the column which has most NA's, it could be a good idea if we remove this column from the original dataset. By dropping this significant variable,we will regain 247 complete observations which will make our further inference more reliable.Before we do the model fitting, we will apply K-function tests.

K-functions in figure 4 confirms our judgement on the clustering of both the cases and the controls. Then whether the cases of non-melanoma skin cancer are completely random among the population? Figure 4 compares the clustering between cases and controls in terms of K-functions(Cao,W., Li,Y., Cheng,J.& Millington,S., 2017). Since the majority of the solid line is in the shadow, see Figure 5, it is might be true that the cancer cases are completely random among population. In consideration of the obvious decreasing trend of the solid line, still an isotropic Gaussian model should be more reliable.

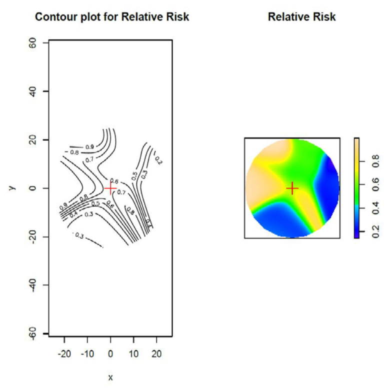

Figure 3: Plots of relative risk of skin cancer

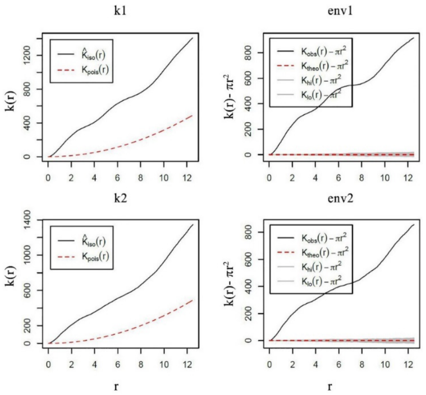

Figure 4: Spatial clustering of cases and controls

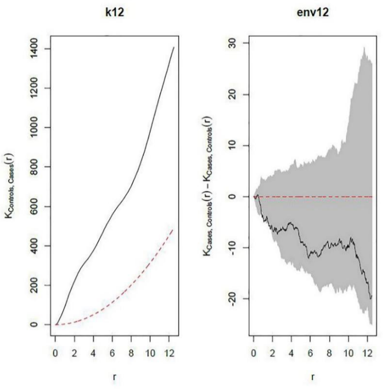

Figure 5: Spatial clustering comparison

5. Conclusion

By fitting the above two models to the dataset, we conclude that urinary arsenic is a factor that associated to the increased risk of non-melanoma skin cancer, also environmental arsenic which based on power station emission is a confounder that related to the increases of this skin cancer risk.

Also, we showed that the distance from the coal-burning power station does not affects the increase of the cancer risk, but it directly affects the environmental arsenic and thus affect the cancer risk but this effect will be easily masked during study.

References:

[1]Ahmed,N., Bodrud-doza,M., Islam,A.R.M.T., Hossain,S., Moniruzzaman,M. et al. . (2019).

[2]Appraising Spatial Variations of As, Fe, Mn and No 3 Contaminations Associated Health Risks of Drinking Water From Surma Basin, Bangladesh. *Chemosphere, 218*.

[3]Bodrud-doza,M., Islam,A.T., Ahmed,F., Das,S., Saha,N. et al. . (2016). Characterization of Groundwater Quality Using Water Evaluation Indices, Multivariate Statistics and Geostatistics in Central Bangladesh. *Water Science, 30*(1).