Modeling Air Quality and Cancer Incidences in Proximity to Hazardous Waste and Incineration Treatment Areas

Miriam Ugarte Querejeta ✉ and Ricardo S. Alonso

[1] International University of La Rioja, Av. de la Paz, 137 26006 Logroño, Spain
miriam.ugarte@estudiante.unir.net,ricardoserafin.alonso@unir.net
[2] BISITE Research Group, University of Salamanca, Edificio Multiusos I+D+i, Calle Espejo 2, 37007 Salamanca, Spain
ralorin@usal.es

Abstract. This study analyzes the impact on human exposure and the air quality in the vicinity to hazardous waste and incineration treatment areas. Such an industry produces pollutant emissions that can be dangerous for the human health and the environment. Thus, various techniques have been studied in order to model the relationship between the proximity to these industrial plants, cancer incidences, and the air quality. On the one hand, logistic regressions were carried out by having the distance as a categorical variable. On the other hand, variable where GAM models were performed. The air quality parameters PM_{10} and NO_2 are higher in proximity to industrial areas according to both techniques, whereas O_3 happens to be lower. Regarding the incidences of cancer, logistic regressions show that the incidences are higher in proximity to certain industrial plants. However, there is no clear conclusion according to the GAM models.

Keywords: air quality · cancer incidence · GAM models · logistic regressions · pollutants.

1 Introduction

Incineration plants often treat and generate hazardous waste and are considered one of the major sources of pollutant emissions such as dioxins and furans [7]. The International Agency for Research on Cancer (IARC) has classified dioxins and furans as carcinogenic hazardous to humans [5] and there are many case studies about the impact of the exposure to hazardous substances on the human health, e.g. Seveso disaster in Italy [18,24,10].

The Basque country is a region in the north of Spain that suffers a high concentration of heavy industry and will be used as a case study. Various statistical studies have been carried out over the past years to analyze the impact on human health and environment in exposure to hazardous waste and incineration plants. Especially, techniques such as linear models, logistic models, and Generalized

2019 ICAI Workshops, pp. 108–122, 2019.

Additive models have been studied in order to assess the air quality and cancer risk in the Basque Country (Spain).

The main objective of this study is to analyze the air quality and cancer incidence in proximity to hazardous waste and incineration treatment areas. On the one hand, the impact of pollutants on human health and environment is described. On the other hand, common data analysis techniques used in other studies have been analyzed. Section 3 Focuses on data analysis techniques of similar studies applied to the air quality and cancer risk. Thus, this study will be focused in logistic regression models and generalized additive models. Section 4 Is the body of the study where it contains the whole data analysis procedure: data collection, data selection and transformation, and data analysis and results. The techniques described on the third section will be applied to analyze the relationship between the cancer incidence and the proximity to hazardous waste sources and also the air quality in areas close to the industrial plants. Finally, the conclusions of the analyses and future work will be described in Section 5.

2 Problem description

Incineration plants are considered one of the main waste treatment systems in many countries [21]. The incineration of hazardous waste produces highly toxic pollutants such as dioxins and dioxin-like compounds that are environmentally persistent and have the ability to bio-accumulate [1]. Waste incineration also can generate heavy metals such as cadmium, mercury and lead and acid gases, among others. Some of these substances are considered carcinogenic to humans according to the International Agency for Research on Cancer (IARC) [5].

2.1 Impact of pollutants on human health and environment

The population is exposed to chemical substances by inhalation of polluted air in its vicinity or consumption of local agricultural products that have been contaminated along the food chain [17].

Furthermore, emissions of hazardous waste processes are transformed into gases, contaminated water, ash, and slag. Substances such as sulfur dioxide (SO_2) and nitrogen dioxide (NO_2) are released into the environment as air, water, and soil [22]. The emissions of these polluting substances may contain adverse impacts on the respiratory health of the human [11,6].

The International legislation (Kiev Protocol on Pollutant Release and Transfer Registers) [2], the European Pollutant Release and Transfer Register Regulation (E-PRTR) [3] and the National Regulations (e.g., Spanish Royal Decree 508/2007 [4]) established a norm in order to regulate the emissions of pollutants and register the inventory of pollutant releases to air, water, and soil.

2.2 Common data analysis techniques

This section explains the techniques and models carried out by previous studies to solve similar statistical problems.

A study carried out in Madrid (Spain) [14] analyzed the cancer mortality in cities close to incinerators and hazardous waste treatment facilities. A higher cancer mortality was observed in the vicinity of industrial facilities and especially to incinerators. Two statistical approaches based on relative log-linear models were used in order to assess the relative risk of mortality: a Bayesian conditional auto regressive model (BYM) and a combined Poisson regression model. Both models showed a significant risk in cancer mortality in the vicinity.

Another study investigated soft tissue sarcomas and non-Hodgkin lymphomas in the vicinity of the Municipal Solid Waste (MSU) incinerator in France that emitted high levels of dioxins. It was observed an increase of 44% increase in soft tissue incidences and an increase of 27% incidences of non-Hodgkin lymphomas within areas near the incinerator. On the one hand, clustering was performed to determine groups within the area. On the other hand, space-time scan statistic method was used to scan multiple data sets in order to look for clusters and evaluate them [23].

Another similar study used Poisson distribution to analyze the incidences of cancer in the vicinity of an incinerator, an oil refinery plant and a waste disposal plant in Rome (Italy). Standardized mortality ratio (SMR) with two 95% confidence intervals were carried out by Poisson distributions [19].

A study analyzed the relationship between the risk of breast cancer and the proximity to industrial plants classified by industrial activities and emissions in Spain. Logistic regressions were used to estimate the Odds Ratio (OR) and 95% confidence intervals of the distance categorized by the proximity to the industrial plant (from 1km to 3km). The results demonstrated a possible increase of risk of breast cancer in women living near certain industrial plants [13].

It exists another study that analyzed the relationship between the air quality particle levels (PM_{10} and $PM_{2.5}$), the meteorological conditions and the traffic. The analysis was carried out by linear regressions and path analysis. The results showed that the weather condition affects the air quality particles PM_{10} and $PM_{2.5}$ in open areas. On the other hand, the results showed that the traffic flow has a direct effect in covered areas. The path analysis was more precise than the linear regression and had a better fit for the study [20].

Zero-inflated regression model was used on a study about the depression influencing factors in a large-scale population survey. The zero-inflated negative binomial was demonstrated to be a good model for the depression factors on a survey type study [25].

3 Data analysis techniques applied to air quality and cancer risk

The aim of this study is to analyze the correlation of different type of tumors and the air quality in proximity to incinerators and hazardous waste treatment plants. The proximity has been considered a categorical variable and also as a continuous numerical variable. Therefore, logistic regressions were applied for the first case study, whereas Generalized Additive Models have been studied on the second case based on recent studies on the field.

3.1 Logistic Regression

An epidemiology study analyzed the association between the air pollution due to heavy industry and lung and respiratory system problems in school children. A cross-sectional study was conducted among children in the vicinity to a heavy industrial area. Linear and logistic regressions were used to carry out the relationship between the air quality parameters ($PM_{2.5}$, NO_X) and lung and respiratory symptoms. The results concluded that the exposure to $PM_{2.5}$ and NO_X was associated with children having lung function problems. The exposure to $PM_{2.5}$ was also associated with children suffering dry cough symptoms [9].

The Seveso disaster study was conducted to investigate the relationship between the air pollution and lung cancer and logistic regressions were used to assess the relative risk of lung cancer in the vicinity of the incinerator [8].

Logistic regressions estimate the parameters of a logistic model. In this study, the cities are classified into two categories according to their distance to the industrial plants bearing the distance of 5km as a threshold [12,14]. Thus, cities with a distance less than or equal to 5 km to the industrial plants have been categorized as *neighbouring cities* and cities with a distance higher than 5 km as *distant cities*. This theorem will perform the comparison of the two observed categories (neighboring cities and distant cities) with respect to a numerical variable (incidences of cancer and air quality).

Logistic regressions have been performed to assess the relationship between the variables. The categorical value is the dependent variable X of the logistic model and will be binomial in this case: *neighbouring city* and *distant city*. The independent variable Y is the predictor and can be a continuous value such as the incidences of cancer and the air quality.

$$Y \sim= B(N, f(X\beta)) \tag{1}$$

3.2 Generalized Additive Model

A study conducted the assessment of cancer mortality in the vicinity of urban solid waste incinerators carried out different statistical techniques to evaluate the association between the risk of cancer and the proximity to incinerators. Techniques such as Poisson regression, general additive models (GAM) and Bayesian hierarchical analysis were used [15].

GAM models were also used to analyze the effects of each pollutant and acute respiratory disease in children. GAM models were performed by applying the quasi-Poisson regression. The study observed that there is an association between exposure to PM_{10}, NO_2 and SO_2 pollutants and a greater number of cases of acute respiratory disease in minors [26].

The Generalized Additive Model (GAM) is an extension of the Generalized Linear Model (GLM). The main idea is to replace the linear component of the model by an additive component. GAM models are formed by the sum of smooth functions (splines) or polynomials $f_i(X_i)$. The purpose is to adjust the smooth non-linear functions into predictor variables X_i to explain the relationship between the dependent variables Y and the predictor variables X_i [16].

$$Y = f_1(X_1) + f_2(X_2) + ... + f_p(X_p) + \epsilon \tag{2}$$

In this case, the dependent variable will be the incidences of cancer or the air quality and the independent variables (predictors) will be the distance to the industrial plant (*dist*), the distance to the closest industrial plant (*dist_min*) and the number of industrial plants per region (n) for each group of tumor, air quality parameter and industrial plant.

4 Experiments and results

R is a programming language widely used for statistical analysis. *R* offers extensive packages and libraries for machine learning and also statistical model packages which results ideal for this study. *Rstudio* is the integrated development environment for *R* and it will be used in this study to carry out the following tasks:

- Data collection
- Data selection and transformation
- Data Analysis and results

4.1 Data collection

The first step is to collect all the data from different sources and to restructure into a common format or data frames in this case. The following data will be retrieved:

Air quality data Air quality data has been obtained through the Air Quality Control Network of the Basque Government. The air quality is daily measured by PM_{10}, NO_2 and O_3 parameters. Parameters are classified into the following categories by the Air Quality Index (IQA): *Very low*, *Low*, *Moderate*, *High* and *Very high*. Thus, the daily Air Quality Index data of each municipality in 2017 has been retrieved.

Industrial plants data The Spanish National Regulation of pollutants contains the inventory of all the industrial plants of hazardous waste treatment and waste incinerators. It reports yearly emissions of all industrial activities to comply with the regulation. Thus, hazardous waste treatment industrial plants of the Basque Country that carried out industrial activities within 2007 and 2016 have been retrieved, in total 109 industrial plants have been studied.

Epidemic data Epidemic data has been obtained by the Public Health Direction of the Basque Government under privacy and data protection terms. The epidemic data is classified by gender, age-range, and city of the diagnosed person. The incidences are divided into 26 groups of different types of tumors according to the International International Classification of Diseases ICD-10.

Population data

4.2 Data Selection and transformation

On one hand, the parameters of interest of each data frame will be selected and common parameters such as the geoposition will be identified in order to be able to join them. On the other hand, one of the main tasks is to calculate the distance between the industrial plants and cities where the incidences of cancer and air quality measurements have been retrieved. The distance will be calculated by the *distGeo* function of *R* as denoted in the following equation:

$$\begin{aligned} setDT(df_1[, dis := distGeo(matrix(c(df_2\$lon.x, df_2\$lat.x), ncol = 2),\\ matrix(c(df_1\$lon.y, df_1\$lat.y), ncol = 2))] \end{aligned} \tag{3}$$

4.3 Data Analysis and results

On the one hand, correlations will be carried out to check if there is any significant relationship between the variables. On the other hand, regressions will be performed to model the relationships. *t-student* and ANOVA metrics are used to assess the results. *t-student* evaluates whether the difference observed between the means of two groups is significant in respect to the hypothesis that has been defined. ANOVA tests evaluate if there is any significant relationship between different populations by comparing the variance of the population.

Logistic regression *t-student* statistic will be used to carry out the analysis in respect to the null hypothesis for its numerical variable.

Having incidences of cancer as a numerical variable:

- **H0** *the mean of incidences of cancer in the neighboring cities is the same as the mean of incidences in the distant cities (null hypothesis).*
- **H1** *the mean of incidences of cancer in the neighboring cities is different than the mean of incidences in the distant cities (alternative hypothesis).*

Having the air quality as a numerical variable:

- **H0** *the mean of the air quality in the neighboring cities is the same as the mean of the air quality in the distant cities (null hypothesis).*
- **H1** *the mean of the air quality in the neighboring cities is different than the mean of the air quality in the distant cities (alternative hypothesis).*

p-value is used to weigh the strength of the evidence. The null hypothesis will be rejected with an interval confidence of 99% if there is a 0.01 probability of rejecting this hypothesis. To completely reject the null hypothesis p-value has to be less than $\alpha = 0.01$ and F value greater than the F-critical.

$$qf(1 - \alpha, df_1, df_2) \tag{4}$$

glm (generalized linear model) functions of the *mgcv* library in *R* are used to calculate logistic regression models.

Incidences of cancer Incidences of cancer are modeled by logistic regressions in order to assess the relationship between the categorical distance and the incidences.

$$glm(km \sim incidences, data = df_{tumour}, family = \text{“}binomial\text{”}) \tag{5}$$

The logistic regression shows that there are more incidences of cancer in the vicinity to industrial plants with a *p-value* < 0.01. Cities that are close to the industrial plant with *NationalID* 3694 have more incidences of cancer of group 20 as shown in Fig. 1:

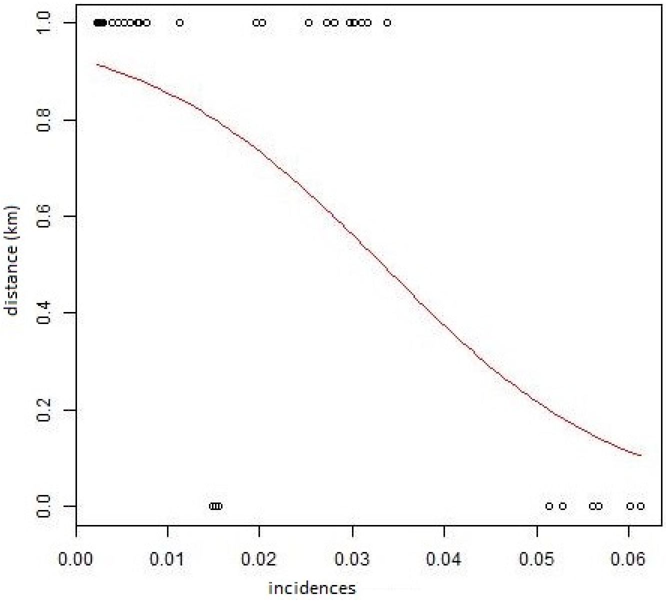

Fig. 1. Logistic regression of cancer incidences, *NationalID*=3694 and group of cancer=20

PM_{10} The air quality index PM_{10} index is modeled by logistic regression to assess the relationship between the categorical distance and the index.

$$glm(km \sim PM_{10}, data = df_{PM_{10}}, family = \text{“}binomial\text{”}) \qquad (6)$$

The logistic regression shows that there is a higher probability of being in a distant city to the industrial plant when the value of the parameter of PM_{10} is low. On the same way, the probability of being in a neighboring city to the industrial plant will be greater when the value of PM_{10} is high as seen in Figure 2.

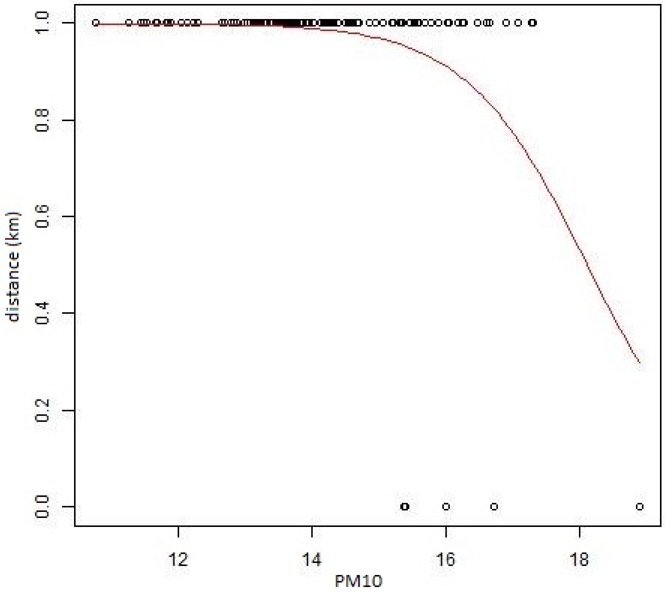

Fig. 2. Logistic regression of PM_{10}, *NationalId*=3698 and *month*=April

NO_2 The air quality index NO_2 index is modeled by logistic regression to assess the relationship between the categorical distance and the index.

$$glm(km \sim NO_2, data = df_{NO_2}, family = \text{“}binomial\text{”}) \qquad (7)$$

There is a higher probability of being in a neighboring city to the industrial plant when the value of the NO_2 parameter is greater as seen in Fig. 3.

O_3 The air quality index O_3 index is modeled by logistic regression in order to assess the relationship between the categorical distance and the index.

$$glm(km \sim O_3, data = df_{O_3}, family = \text{“}binomial\text{”}) \qquad (8)$$

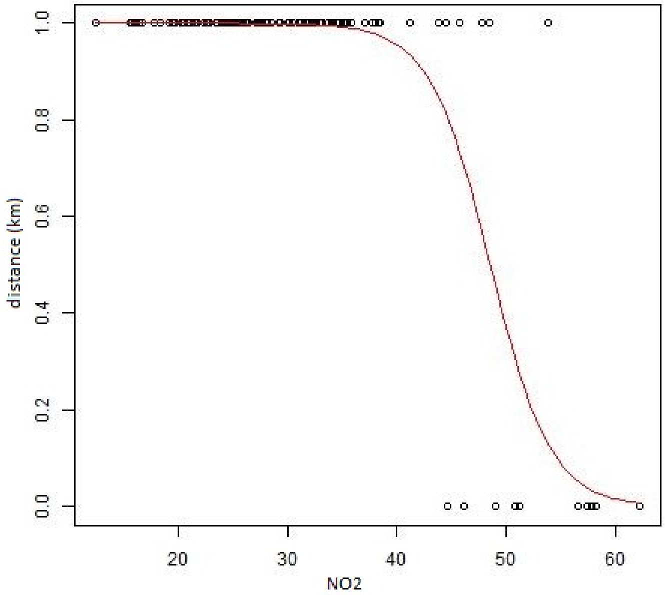

Fig. 3. Logistic regression of NO_2, *NationalId*=3702 and *month*=April

There is a higher probability of being in a neighboring city to the industrial plant when the value of the O_3 parameter is low as seen in Fig.4.

Generalized Additive Model The linear predictor is not forced to be linear in this case and it is constructed by the sum of smooth functions called splines. The variable can be continuous, categorical, linear, data series, etc. Thus, *gam* functions of the *mgcv* library in *R* are used to calculate the generalized additive models.

Incidences of cancer Incidences of cancer are modeled by GAM so as to assess the relationship between the distance and the incidences.

$$gam(incidences \sim s(n, k = 20) + s(dist_{min}, k = 20) + s(dist, k = 20)) \qquad (9)$$

The GAM model varies with the industrial plant and the group of tumor and there is no clear pattern between the incidences and the distance to the industrial plants. The model can not be generalized as there might be areas affected by other factors as seen in Fig. 5.

PM_{10} The air quality index PM_{10} is modeled by GAM to assess the relationship between the distance and the index.

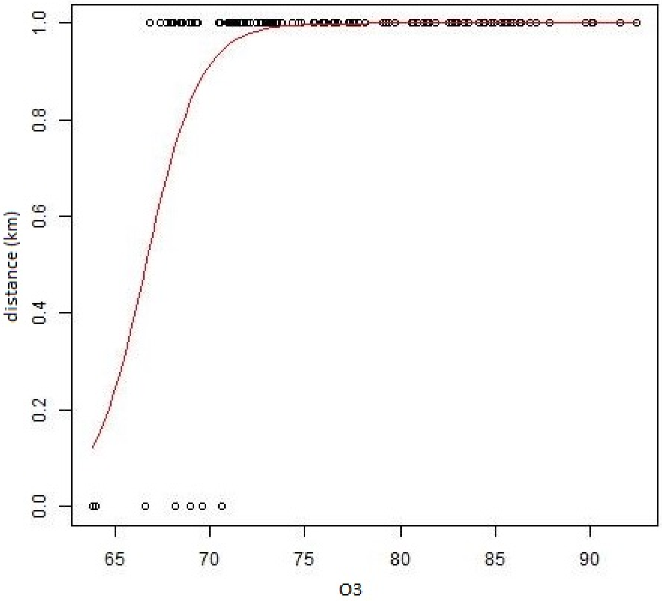

Fig. 4. Logistic regression of O_3, *NationalId*=4682 and *month*=August

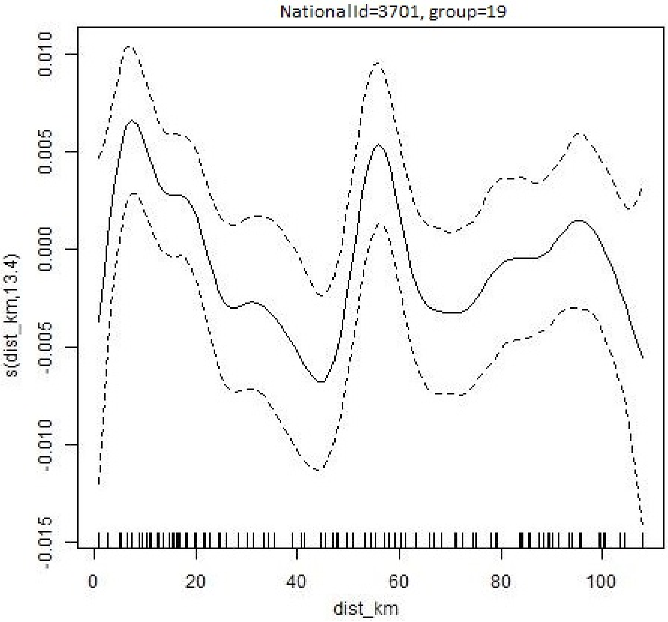

Fig. 5. GAM model of cancer incidences, *NationalID*=3701 and group of *cancer*=16

$$gam(PM_{10} \sim s(n, k = 20) + s(dist_{min}, k = 20) + s(dist, k = 20)) \quad (10)$$

Regarding the GAM model, the value of the parameter PM_{10} is higher (worse AQI) in the vicinity of the industrial plant and it decreases and improves with the distance as described in Fig. 6.

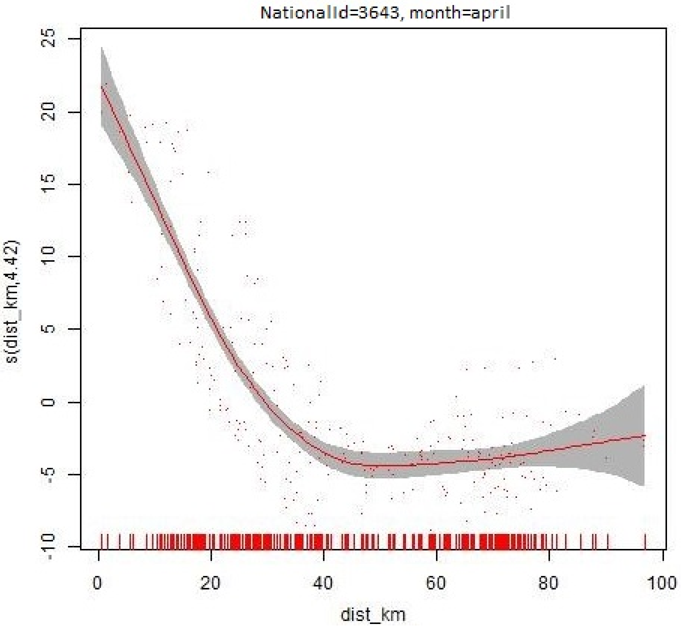

Fig. 6. GAM model of PM_{10}, *NationalId*=3643 and *month*=April

NO_2 The air quality index NO_2 is modeled by GAM in order to assess the relationship between the distance and the index.

$$gam(NO_2 \sim s(n, k = 20) + s(dist_{min}, k = 20) + s(dist, k = 20)) \quad (11)$$

The parameter NO_2 is greater in the vicinity of the industrial plant and it decreases with the distance according to the GAM model as seen in Fig. 7.

O_3 The air quality index O_3 is modeled by GAM to assess the relationship between the distance and the index.

$$gam(O_3 \sim s(n, k = 20) + s(dist_{min}, k = 20) + s(dist, k = 20)) \quad (12)$$

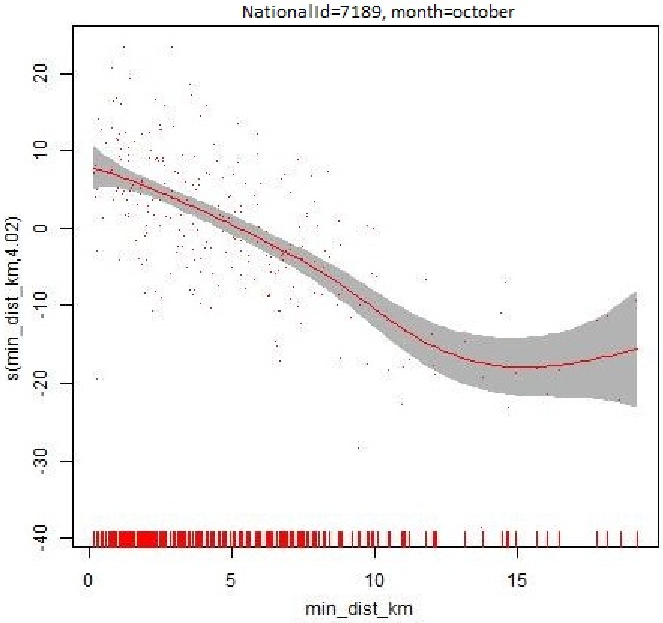

Fig. 7. GAM model of NO_2, *NationalId*=7189 and *month*=October

The following Figure shows the GAM model for the *NationalId* 5916 and the parameter O_3 in September. The value of O_3 is increasing linearly until being 60 km far from the industrial plant. At this point, the increase of O_3 parameter gets smaller as described in Fig. 8.

5 Conclusions and Future Work

Logistic regression model results show a significant association ($p < 0.01$) between cancer incidences and the proximity to industrial plants of hazardous waste treatment. Logistic regression models also show that PM_{10} and NO_2 values are higher in the proximity to certain industrial plants with $p < 0.01$. However, O_3 levels happen to be lower.

GAM model does not fit well to represent the association between incidences of cancer as the model changes by the type of tumor and the industrial plant. The results can not be generalized by GAM models as certain tumors may be biased by other factors that have not taken into account. For example, lung cancer incidences could be biased by tobacco consumption. However, GAM models have demonstrated to be a good fit to represent the relationship between air quality and the proximity to industrial plants. PM_{10} and NO_2 levels are higher in the vicinity to these industrial plants whereas O_3 levels are lower. GAM models seem to fit better than logistic regressions to model the behavior of the air quality.

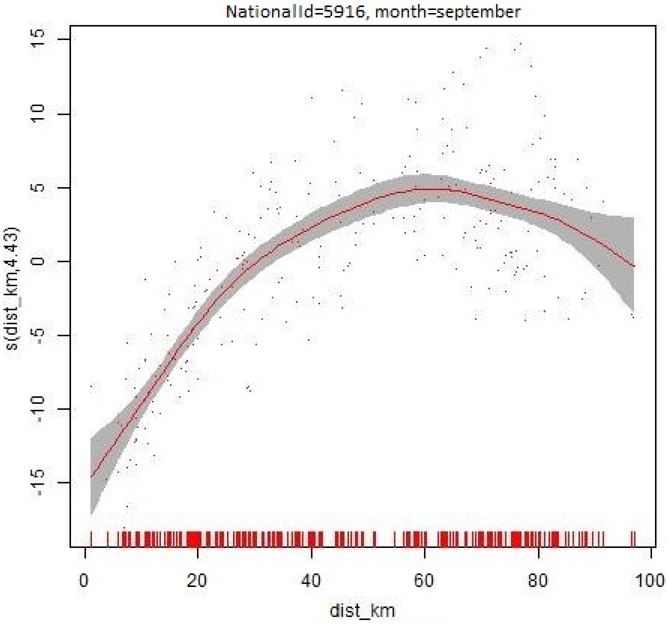

Fig. 8. GAM model of O_3, *NationalId*=5916 and *month*=September

Atmospheric emissions may be conditioned by meteorological variables such as precipitation, temperature, wind, etc. and thus, it would be interesting to study the impact of weather conditions as a future work. Also, industrial plants could be classified by the activities carried out and the emitted substances in order to go deeper in the analysis and find a relationship with the substances that are emitted.

Acknowledgements This work has been supported by the International University of La Rioja (Spain).

References

1. The National Academies, Waste Incineration and Public Health (2000)
2. United Nations, Protocol On Strategic environmental assessment to the convention on environmental impact assessment in a transboundary context. Economic and Social Council for Europe, Kiev (May 2003)
3. Regulation (EC) No 166/2006 of the European Parliament and of the Council. Official Journal of the European Union (Jan 2006)
4. Ministerio de medio ambiente, Real Decreto 508/2007. BOE num 96 (BOE-A-2007-8351), 17686–17703 (Apr 2007)
5. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, A Review of Human Carcinogens. International Agency for Research on Cancer, Lyon (France) **100** (2011)

6. Effects of VOCs on Human Health, Air Pollution and Control. Energy, Environment, and Sustainability, Springer pp. 119–142 (2017)
7. Plan de prevención y gestión de residuos de la CAPV/2020. Departamento de Medio Ambiente Planificación Territorial y Vivienda, Servicio Central de Publicaciones del Gobierno Vasco, Donostia-San Sebastiánenda, (2018)
8. Barbone, F., Bovenzi, M., Cavallieri, F., Stanta, G.: Air Pollution and Lung Cancer in Trieste, Italy. American Journal of Emidemiology **141**(12) (1995)
9. Bergstra, A.D., Brunekreef, B., Burdorf, A.: The effect of industry-related air pollution on lung function and respiratory symptoms in school children. Environmental Health (2018)
10. Bertazzi, P., Pesatori, A.C., Consonni, D., Tironi, A., Landi, M., Zocchetti, C.: Cancer Incidence in a Population Accidentally Exposed to 2,3,7,8-Tetrachlorodibenzopara-Dioxin. Epidemology, Lippincott Williams & Wilkins **4**(5), 398–406 (1993)
11. Chen, T., Gokhale, J., Shofer, S., Kuschner, W.G.: Outdoor air pollution: nitrogen dioxide, sulfur dioxide, and carbon monoxide health effects. The Americal Journal of the Medical Sciences **3**(4), 249–256 (2007)
12. Federico, M., Pirani, M., Rashid, I., Caranci, N., Cirilli, C.: Cancer incidence in people with residential exposure to a municipal waste incinerator: an ecological study in Modena (Italy), 1991-2005. Waste Management **30**(7), 1362–70 (2010)
13. García, J., Lope, V., Pérez-Gómez, B., Molina, A.J., Tardón, A., Díaz Santos, M.A., Ardanaz, E., O'Callaghan-Gordo, C., Altzibar, J.M., Gómez-Acebo, I., Moreno, V., Peiró, R., Marcos-Gragera, R., Kogevinas, M.: Risk of breast cancer and residential proximity to industrial installations: New findings from a multicase-control study (MCC-Spain). Environmental Pollution **237**, 559–568 (2018)
14. García-Pérez, J., Fernández-Navarro, P., Castellóa, A., López-Cima, M., Ramis, R., Boldo, E., López-Abente, G.: Cancer mortality in towns in the vicinity of incinerators and installations for the recovery or disposal of hazardous waste. Environmental International **51**, 31–44 (2013)
15. Goria, S., Daniau, C., Crouy-Chanel, P.D., Empereur-Bissonnet, P., Fabre, P., Colonna, M., Doboudin, C., Viel, J.F., Richardson, S.: Risk of cancer in the vicinity of municipal solid waste incinerators: importance of using a flexible modelling strategy **8**(31) (2009)
16. Harezlak, J., Ruppert, D., Wand, M.: Generalized Additive Models, Semiparametric Regression with R. Springer pp. 71–128 (2018)
17. Hoogenboom, R.: Incidents with dioxins and dioxin-like PCBs in the food chain. ECVPH Food safety assurance **7**, 503–528 (2018)
18. Marcella, W., Paolo, M., Steven, S., Larry, N., Paolo, B., Brenda, E.: Dioxin Exposure and Cancer Risk in the Seveso Women's Health Study. Environmental Health Perspectives pp. 1700–1705 (2011)
19. Michelozzi, P., Fusco, D., Forastiere, F., Ancona, C., Dell'Orco, V., Perucci, c.A.: Small area study of mortality among people living near multiple sources of air pollution. Occupational and Environmental Medicine **55**(9), 611–615 (1998)
20. Sahanavin, N., Prueksasit, T., Tantrakarnapa, K.: Relationship between pm10 and pm2.5 levels in high-traffic area determined using path analysis and linear regression. Elsevier pp. 105–114 (2018)
21. Scarlat, N., Fahl, F., Dallemand, J.: Waste and Biomass Valorization, Status and Opportunities for Energy Recovery from Municipal Solid Waste in Europe. Springer **10**(9), 2425—-2444 (2019)
22. Tiwary, A., Williams, I.: Air Pollution Measurement, Modelling and Mitigation, Fourth Edition. CRC Press, Boca Raton (2018)

23. Viell, J.F., Arveux, P., Baverel, J., Cahn, J.Y.: Soft-Tissue Sarcoma and Non-Hodgkin's Lymphoma Clusters around a Municipal Solid Waste Incinerator with High Dioxin Emission Levels. American Journal of Epidemiology **152**(1) (2000)
24. Warner, M., Mocarelli, P., Samuels, S., Needham, L., Brambilla, P., Eskenazi, B.: Dioxin Exposure and Cancer Risk in the Seveso Women's Health Study. Environmental Health Perspectives pp. 1700–1705 (2011)
25. Xu, T., Zhu, G abd Han, S.: Study of depression influencing factors with zero-inflated regression models in a large-scale population survey. BMJ Open **7**(11) (2017)
26. Zhu, L., Ge, X., Chen, Y., Zeng, X., Pan, W., Zhang, X., Ben, S., Yuan, Q., Xin, J., Shao, W., Ge, Y., Wu, D., Han, Z., Zhang, Z., Chu, H., Wang, M.: Short-term effects of ambient air pollution and childhood lower respiratory diseases. Scientific Reports **7** (2017)