A modelling framework for developing early warning systems of COPD emergency admissions

Olatunji Johnson[a,*], Tim Gatheral[b], Jo Knight[a], Emanuele Giorgi[a]

[a] *CHICAS Research Group, Lancaster Medical School, Lancaster University, Bailrigg, Lancaster, UK*
[b] *Respiratory Medicine, Royal Lancaster Infirmary, Lancaster, UK*

Abstract

Chronic Obstructive Pulmonary Disease (COPD) is one of the leading causes of mortality worldwide and is a major contributor to the number of emergency admissions in the UK. We introduce a modelling framework for the development of early warning systems for COPD emergency admissions. We analyse the number of COPD emergency admissions using a Poisson generalized linear mixed model. We group risk factors into three main groups, namely pollution, weather and deprivation. We then carry out variable selection within each of the three domains of COPD risk. Based on a threshold of incidence rate, we then identify the model giving the highest sensitivity and specificity through the use of exceedance probabilities. The developed modelling framework provides a principled likelihood-based approach for detecting the exceedance of thresholds in COPD emergency admissions. Our results indicate that socio-economic risk factors are key to enhance the predictive power of the model.

Keywords: COPD; early warning system; exceedance probabilities; generalised linear mixed model; spatio-temporal models;

1. Introduction

Chronic Obstructive Pulmonary Disease (COPD) is one of the leading causes of mortality worldwide (Mathers & Loncar, 2006; Hasegawa et al., 2014) with an

[*] Corresponding author
Email address: o.johnson@lancaster.ac.uk (Olatunji Johnson)

estimated 3 million deaths in 2015, corresponding to 5% of all deaths globally (World Health Organisation, 2016). Acute exacerbations are a major contributor to the number of emergency admissions and hospitalizations Tian et al. (2012), especially during the winter months as a result of the increase in respiratory viral infections. Most of the current research has been focused on understanding the risk factors associated with COPD exacerbation (Osman et al., 2017; Bahadori & FitzGerald, 2007; Chan et al., 2011). While the majority of exacerbations are caused by infectious agents, especially rhinoviruses Wedzicha (2004), there has been evidence from previous studies that biological, environmental and socio-economic factors can also trigger COPD emergency hospitalisation (Hemming et al., 2009). Hemming et al. (2009) developed a Bayesian network approach in order to identify factors that can help predict COPD admissions in the UK and found a combination of environmental, socio-economic and health-related variables to be useful predictors. These included weather type (classified as sunny, cloudy, rainy, windy and snowy) temperature, outdoor air pollution, gas emissions, urbanisation, smoking, population age, environmental tobacco smoke, indoor air pollution (housing condition), income and education, infection load, number of previous admission and severity of the disease. However, most studies have examined these factors separately and only a few have assessed their joint contribution to COPD risk.

Predictive models have been developed in several studies to identify patients at high risk of COPD exacerbations (Billings et al., 2006; Yii et al., 2019; Urwyler et al., 2019; Samp et al., 2018) which add significant cost to the patients care. Hence, being able to accurately predict their occurrence can be especially useful in order to reduce avoidable COPD emergency admissions by targeting patients in most need. In order to develop a robust and scalable predictive model for COPD emergency admissions, the availability of comprehensive health records of patients is essential so as to ensure its reliability. Predictive power can also be further enhanced by incorporating risk factors concerning the lifestyle behaviour (e.g. smoking status), income, exposure to pollutants and other individual traits. However, such detailed information may not be readily available to researchers due to confidentiality issues or because it has not been collected. Notwithstanding, statistical modelling provides solutions that can be used to alleviate this issue. For example, generalized linear mixed models (GLMMs) (Breslow & Clayton, 1993) are an extension of the classical generalized linear modelling framework that allows to account for the unavailability of risk factors through the use of so-called random effects. However, the full potential offered by this modelling framework has not been fully exploited in the analysis of COPD data. In this paper, we aim to address this gap.

While some analyses on COPD emergency admissions have focused on individual or patient level where biological markers (e.g. forced expiratory volume in 1 seconds (Wei et al., 2018) and blood eosinophil level (Bélanger et al., 2018)) were used to predict the risk of an emergency admission, here we focus our attention on studies that were concerned with understanding the geographical variation of COPD risk at population-level. Niyonsenga et al. (2018) model the prevalence of COPD and asthma over census units in the western area of Adelaide, South Australia, and assessed the spatial clustering of cases using the local Getis-Ord's Gi indices (Anselin, 1995). Kauhl et al. (2018) analyse how the prevalence of COPD varies across northeastern Germany and identify risk factors including proportions of insurants aged above 65, proportions of insurants with migration background, household size and area deprivation as statistically significant predictors for COPD. Holt et al. (2011) were the first to characterise geographic variations in COPD hospitalization across Health Service Areas (HSAs) and at state level across the United States. They found distinct geographical pattern in COPD hospitalisation rate in the HSA and state level, suggesting that different risk factors could be operating at different spatial scales. In another study conducted in Taiwan, Chan et al. (2014) analyse the spatio-temporal distribution of COPD mortality over a 9 year period, from 1999 to 2007. They found that smoking rate, the percentage of aborigines within a district, PM_{10}, altitude and density of healthcare facilities were significantly associated with COPD mortality.

Most spatio-temporal analyses on COPD have used conditional autoregressive models (CAR) (Besag et al., 1991) to carry out spatial smoothing of COPD risk but did not attempt any forecasting. CAR models are formulated by defining a correlation structure between neighbouring areal units (e.g. districts or regions). In addition, all of these studies (Kauhl et al., 2018; Holt et al., 2011; Chan et al., 2014) have focused their efforts in predicting mean level of risks. In this paper, we argue that statistical modelling should, instead, aim to predict the exceedance of clinically relevant thresholds beyond which COPD risk is of public health concern.

In our analysis of COPD admissions, we pursue two specific objectives: 1) to assess the relative contribution of socio-economic and environmental variables for forecasting COPD emergency admissions; 2) to develop a reliable surveillance system that triggers an alarm whenever COPD emergency admissions signal the likely exceedance of predefined incidence thresholds. To the best of our knowledge, this is the first study that attempts to achieve these objectives using state-of-the-art spatio-temporal statistical methods for the analysis of data on COPD emergency admissions.

2. Methods

2.1. COPD admission data

Using the International Classification of Diseases (ICD) code (10th revision)49 , J44 for COPD, we extracted monthly counts of COPD emergency admissions for patients above 19 years living in the LA postcode area, covering parts of South Cumbria and North Lancashire in England (see Figure 1). The total population of the study region was 272,520 based on the 2011 census. The data cover the period from 1 April 2012 to 30 March 2018. To protect confidentiality and anonymity of the patients, spatial information on their place of residence was provided at the Lower Super Output Area (LSOA).

The COPD emergency admission data was provided by Morecambe Bay NHS Foundation Trust (UHMBT). The study received approval from the research and development department of UHMBT.

2.1.1. Environmental variables

We obtained monthly weather data for 2012-2018 including monthly relative humidity, number of days of ground frost and temperature from the UK Met Office, freely available from the Centre for Environmental Data Analysis (`http://data.ceda.ac.uk/`). The spatial resolution of the weather raster files is $1 \times 1\text{km}^2$ across the UK. We also obtained yearly pollution data including Particulate Matter less than 10 m in diameter (PM_{10}), Sulphur Dioxide (SO_2) and Nitrogen Dioxide (NO_2), available from the Department of Environmental Food and Rural Affairs (DEFRA) (`https://uk-air.defra.gov.uk/data/pcm-data`). The estimate of the pollutants are provided at $1 \times 1\text{km}^2$resolution over the entire Great Britain. For our analysis, we computed the population weighted average of all the available raster data over the LSOAs shown in Figure 1.

2.1.2. Socio-economic variables

We obtained the index of multiple deprivation (IMD) created by the Department for Communities and Local Government in order to account for socio-economic heterogeneities across LSOAs. The IMD combines seven domains which relate to income deprivation, employment deprivation, health deprivation and disability, education skills and training deprivation, barriers to housing and services, living environment deprivation, and crime. The IMD is available as either a score, decile or rank. In this study, we used the IMD score for 2015, the most recent release. Larger values of the score correspond to a higher level of the domain deprivation.

2.1.3. Population data

We obtained the yearly population data per LSOA from the Office of National Statistics (ONS), UK. ONS updates their population estimates yearly based on migration data and any other physical adjustments (Office for National Statistics, 2018). The average population of LSOAs in England and Wales according to the census data in 2011 was 1,614 with 95% of LSOAs having a

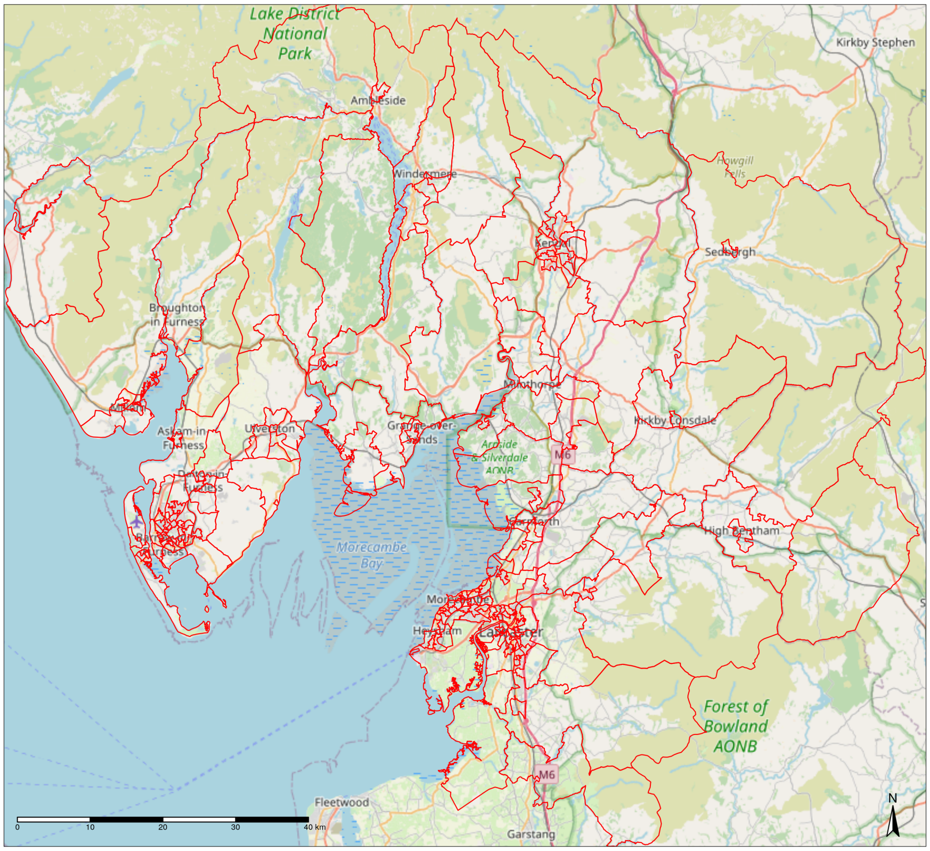

Figure 1: Map of South Cumbria and North Lancashire showing the boundaries (blue lines) of 209 Lower Super Output Areas (LOSAs).

population of between 1,157 and 2,354. We stratify the population data into age and sex using age group $20-29$, $30-39$, $40-49$, $50-59$, $60-69$, $70-79$, $80-89$ and 90+ and sex group male and female.

2.2. Statistical modelling and assessment of residual spatio-temporal correlation

Let Y_{it} denote the monthly COPD emergency admission count at LSOA i and month t. We then assume that the Y_{it}, conditionally on a random effect Z_{it}, follow a Poisson distribution with mean $\mu_{it} = E_{it}\lambda_{it}$, where E_{it} denotes the expected counts at LSOA i and month t and λ_{it} represents the monthly relative risk of COPD emergency admission at a given LSOA.

We compute E_{it} using indirect standardization for age and sex. Let $m = 16$ denote the number of age and sex strata and let n_{itj} and y_{itj} is the total population and the number of COPD emergency admission, respectively at LSOA i, month t and in stratum j. Then, the expected counts is computed as follows:

$$E_{it} = \sum_{j=1}^{m} \hat{r}_j^{(s)} n_{itj},$$

where $\hat{r}_j^{(s)} = \sum_{it} y_{itj} / \sum_{it} n_{itj}$ is the disease rate in stratum j in the standard population, such that $\bar{r} = \sum_{j=1}^{m} \hat{r}_j^{(s)}/m$.

We define the log-linear model for the mean number of cases as

$$\mu_{it} = E_{it} \exp\{d_{it}^{\top}\beta + Z_{it}\}, \tag{1}$$

where d_{it} is a vector of covariates with associated regression coefficients β. Finally, we assume that the Z_{it} are independent and identically distributed Gaussian variables with mean zero and variance σ^2. In order to build our regression model, we select predictors within three domains that are known to affect COPD admissions: weather, pollution and deprivation. The variables that we consider within each of these domains are listed in Table 1. As the variables within each group are highly collinear, our goal is to select the best predictor from each of the three groups.

In order to carry out the selection of the best predictors, we split the dataset into training and test sets, with the former covering the months from April 2012 to March 2017 and the latter from April 2017 to March 2018. The rationale for the chosen test and training sets is that we aim to develop an early warning system that can better capture temporal features of the latest reported admissions. We then fit 63 models obtained by combining one predictor from each domain of Table 1 and, for each of those, we compute the bias, and root-mean-square-error (RMSE) for the predicted COPD admissions cases using the test set. Let ψ_{it} denote the observed incidence rate per 1000 population and $\hat{\psi}_{it}$ denote the

Table 1: Predictors for chronic obstructive pulmonary disease (COPD) and their corresponding domain.

Domain	**Variables**
Weather	Minimum temperature; relative humidity; and number of days of ground frost.
Pollution	PM_{10} SO_2; and NO_2. All in micrograms per cubic metre (μgm^{-3})
Deprivation	Income deprivation; employment deprivation; health deprivation and disability; education skills and training deprivation; barriers to housing and services; living environment deprivation; and crime deprivation.

predicted incidence rate per 1000 population for LSOA t at time t for the test set, hence,

$$\text{Bias} = \frac{1}{IT}\sum_{i=1}^{I}\sum_{t=1}^{T}(\hat{\psi}_{it} - \psi_{it}),$$

$$\text{RMSE} = \sqrt{\frac{1}{IT}\sum_{i=1}^{I}\sum_{t=1}^{T}(\hat{\psi}_{it} - \psi_{it})^2},$$

. For different combination of the set of predictors, we provide a rank for each metric according to their performance. Then compute the cumulative rank across the metrics and choose the combination of predictors with the lowest cumulative rank as the best set.

From the mixed model with the best set of predictors identified through the procedure outlined above, we assess whether the random effects Z_{it} show evidence of residual spatio-temporal correlation. To this end, we compute the empirical spatio-temporal variogram (ESTV) for the estimates of Z_{it}, using the centroid of each LSOA in order to quantify the geographical proximity between LSOAs. Let $\hat{Z}(x_i, t_i)$ denote the estimate of Z_{it} from model (1) associated with the centroid x_i at time t_i. Let $n(u, v)$ denote the pairs (i, j) such that $\|x_i - x_j\| = u$, where $\| \cdot \|$ is the Euclidean distance, and $|t_i - t_j| = v$. The expression of the ESTV is

$$\hat{\gamma}(u,v) = \frac{1}{2|n(u,v)|} \sum_{(i,j)\in n(u,v)} \{\hat{Z}(x_i,t_i) - \hat{Z}(x_j,t_j)\}^2,$$

where $|n(u,v)|$ is the number of pairs set.

We used Monte Carlo methods to construct a 95% confidence intervals for $\hat{\gamma}(u,v)$ under the assumption of absence of spatial correlation. We then proceed through the following iterative steps:

1. permute the order of $\hat{Z}(x_i,t_i)$, while holding (x_i,t_i) fixed;
2. compute the empirical variogram for $\hat{Z}(x_i,t_i)$;
3. repeat step 1 and 2 for a large number of times, say B times; and
4. use the resulting B empirical variogram to generate 95% tolerance interval at each of the predefined distance bins.

If $\hat{\gamma}(u,v)$ lies outside these intervals, then we conclude that the $Z(x_i,t_i)$ shows an evidence of residual spatio-temporal variation. Conversely, if $\hat{\gamma}(u,v)$ lies inside, we conclude that the data do not show evidence against the model in (1) which assumes independence between the counts Y_{it} after removing the effects of the covariates d_{it}.

To quantify the relative contribution of each predictor in the model, we compute the relative variance reduction (RVR) defined as

$$RVR = \frac{\sigma^2_{-j} - \sigma^2}{\sigma^2_{-j}}$$

where σ^2 the variance of the Z_{it} from the final model and σ^2_{-j} is the variance of the Z_{it} when the $j^{-\text{th}}$ predictor is excluded from the final model.

2.2.1. An early warning system based on exceedance probabilities

We aim to develop an early warning system that triggers an alarm if the COPD emergency admissions rate for a given LSOA exceeds a policy relevant threshold, denoted by l.

Using the best model obtained from the previous stage of the analysis, we predicted the expected number of cases as, $\hat{\mu}_{it} = E[\hat{Y}_{it}|\hat{Z}_{it}] = E_{it}\exp\{d_{it}^{\top}\hat{\beta} +$

$\hat{Z}_{it}\}$, such that the expected incidence rate per 1,000 population is given by $\hat{\psi}_{it} = 1,000 \times \bar{r} \times \hat{\mu}_{it}/E_{it}$. We then compute the exceedance probability (EP), i.e. the predictive probability that $\hat{\psi}_{it}$ exceeds a predefined threshold l, formally expressed as

$$EP_{it} = \Pr\left(\hat{\psi}_{it} > l \,\middle|\, y_{it}\right). \tag{2}$$

Values of EP close to one indicate that incidence rate per thousand is highly likely to be above l, while the values of EP close to zero indicate that incidence rate per thousand is highly likely to be below l. Finally, values of EP around 0.5 indicate that incidence rate per thousand are equally likely to be above or below l, thus implying a scenario with highest uncertainty.

For a given LSOA and month, an alarm is then triggered whenever the EP exceeds a value, say p. To identify an optimal value of p, we maximise the *sensitivity* and *specificity* of the early warning system using the test set from April 2017 to March 2018. Sensitivity is computed as the proportion of districts whose true incidence rate per thousand is above the threshold l and are correctly classified based on EP $> p$; the specificity is the proportion of districts whose true incidence rate per thousand is below the threshold l and are correctly classified based on EP $< p$.

We also compare the use of exceedance probabilities with a naive approach which triggers an alarm for an LSOA if the predictive mean for the incidence rate per thousand exceeds l, i.e.

$$E\left[\hat{\psi}_{it} \,\middle|\, y_{it}\right] > l. \tag{3}$$

Note that this approach, unlike (2), ignores the dispersal of the predictive distribution $\hat{\psi}_{it}$ hence yielding a lower sensitivity or specificity for the early warning system.

3. Result

3.1. Descriptive Analysis

The age distribution of the COPD admissions is shown in Figure 2. We observe the largest number of admissions for the age group 70-79. Also, more females were admitted than males.

As expected, the empirical pattern of monthly counts of COPD emergency admission showed a seasonal pattern with the highest peaks found in the winter period each year, around December and January (Figure 3), and lowest number of admissions in September. It is well established that COPD patients suffer from increased exacerbation during cold weather (Donaldson et al., 1999). Our model captures the seasonal variations through the variables falling under the weather domain.

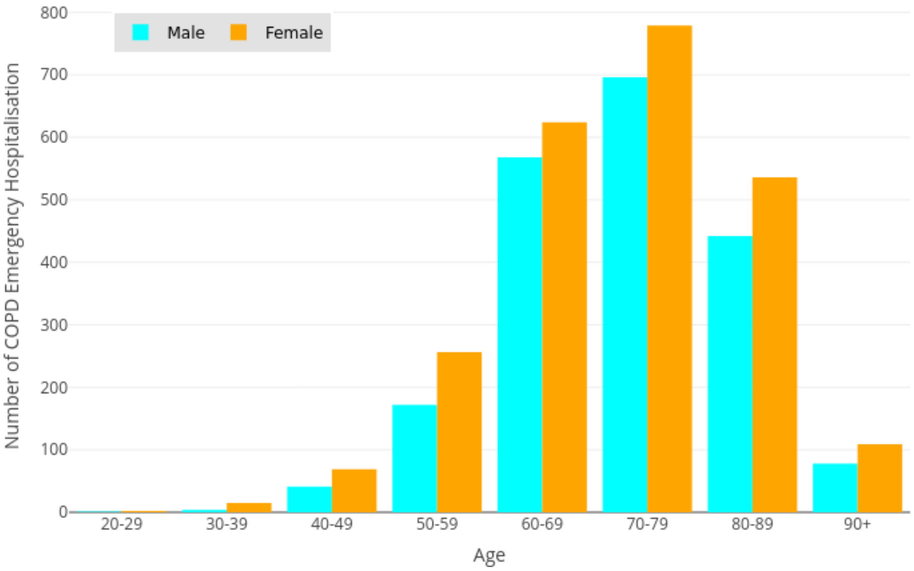

Figure 2: Count of COPD emergency admission, by age group and sex, in South Cumbria and North Lancashire, 2012-2018

3.2. Spatio-temporal Analysis

By applying the variables selection procedure described in Section 2.2, our final set of predictors consists of minimum temperature, PM_{10}, income deprivation (see Table A.3 and A.4 in supplementary material). This set of predictors has rank first in terms of cumulative rank among other candidate set of predictors.

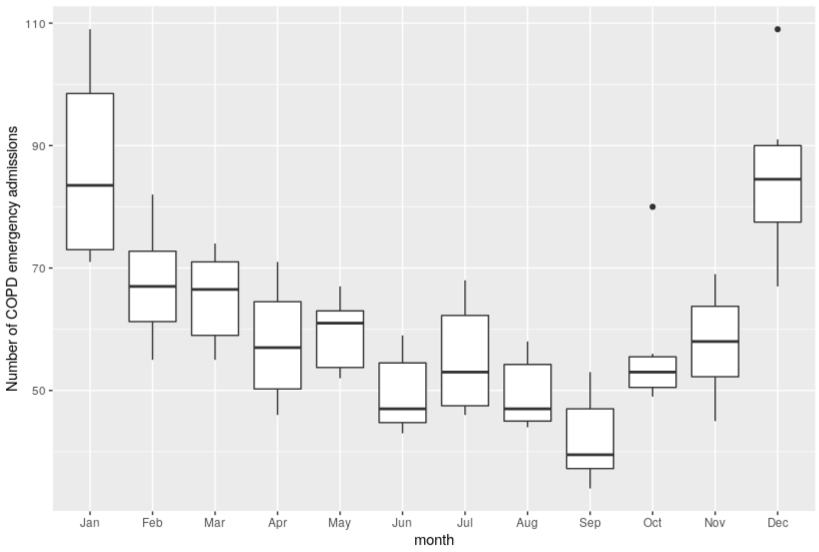

Figure 3: Boxplot showing seasonal variation in the monthly count of COPD emergency admission in South Cumbria and North Lancashire, 2012-2018.

Table 2 shows the relative variance reduction (RVR) of each predictor in the model. We find that, overall, the selected predictor variables explain about 17% of the variability in the residual random effects. Among these variables, income deprivation attained the largest RVR of about 12%.

In order to test whether the predictors included in this model can capture all the spatio-temporal correlation in the data, we applied the Monte Carlo procedure of Section 2.2 based on the spatio-temporal variogram for both an intercept-only model, that excludes all of the predictors of the final model (Figure 4), and the final model (Figure 5). The empirical variogram is based on spatio-temporal bins that span in space and time. A comparison between Figures 4 and 5 indicates that the predictors used in the final model allowed us to capture most of the residual spatio-temporal correlation in COPD emergency admissions. For this reason, we deemed the model with unstructured Z_{it} to be a satisfactory fit to the data.

We then predict the incidence rates per thousand of COPD emergency admission for April 2017 - March 2018 and classify each LSOA as being above or below an incidence rate per thousand threshold l which we set to 12 per 1,000, a choice which was agreed in consultation with expert clinicians. Based on this threshold, the value of EP value that maximizes the sensitivity and specificity of the early warning system was 0.85, yielding a 72% sensitivity and a 70% specificity. The area under the curve (AUC) was about 78% (Figure 6 left panel), indicating a satisfactory predictive performance. The approach which classifies LSOAs based on the predictive mean yields a 70% sensitivity and a 58% specificity. Because of the very low specificity of this second approach which leads to an unacceptable high number of false alarms, the use of exceedance probabilities is our preferred classification procedure.

Figure 7 shows the LSOA that were correctly and incorrectly classified based on the exceedance probabilities. The selected months were chosen to show the performance of the model in each of the four seasons of the test set. LSOAs that are incorrectly classified corresponds to false alarms (purple and red LSOAs). In October 2017, we observe the largest number of COPD emergencies that are detected correctly by the model corresponding to 160 out of 209 LSOAs (blue and orange LSOAs).

Table 2: The table showing the relative variance reduced by the predictors.

Predictors	**RVR (%)**
Minimum temperature	0.30
PM_{10}	0.18
Income deprivation	11.51
All predictors	17.01

4. Discussion

We have proposed a modelling framework that allows the development of an early warning system for chronic obstructive pulmonary disease (COPD)

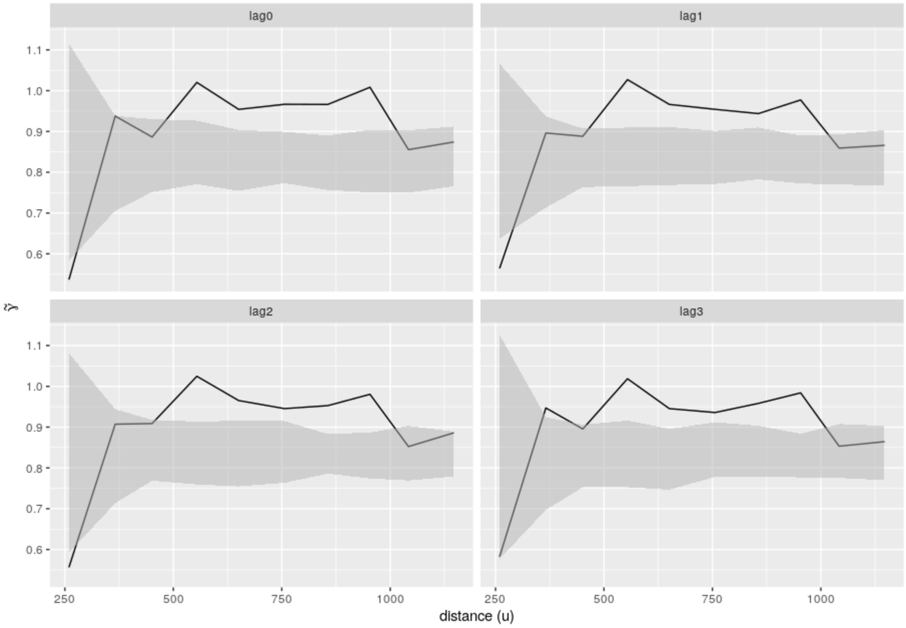

Figure 4: Spatio-temporal variogram of the residual from an intercept only model.

admissions and have applied this to a case study in South Cumbria and North Lancashire districts of Northwest England.

We have used our novel approach to identify emergencies in the number of COPD admissions which are triggered whenever a predefined level of incidence rate is exceeded. We have argued that, in this context, the development of statistical models should aim to quantify the risk for the occurrence of a major public health emergency through the use of exccedance probabilities (EPs). Unlike other commonly used indices of predictive performance, such as mean square errors, EPs are easier to interpret and more directly address the public health question raised by this study. This contrasts with the current prevailing approach, where models are exclusively developed in order to optimize predictions for average levels of incidence.

Our best fitting model was identified by optimizing the exceedance probability threshold for classification of LSOAs in order to maximize both sensitivity and specificity. However, if the prioritization of LSOAs with incidence rate

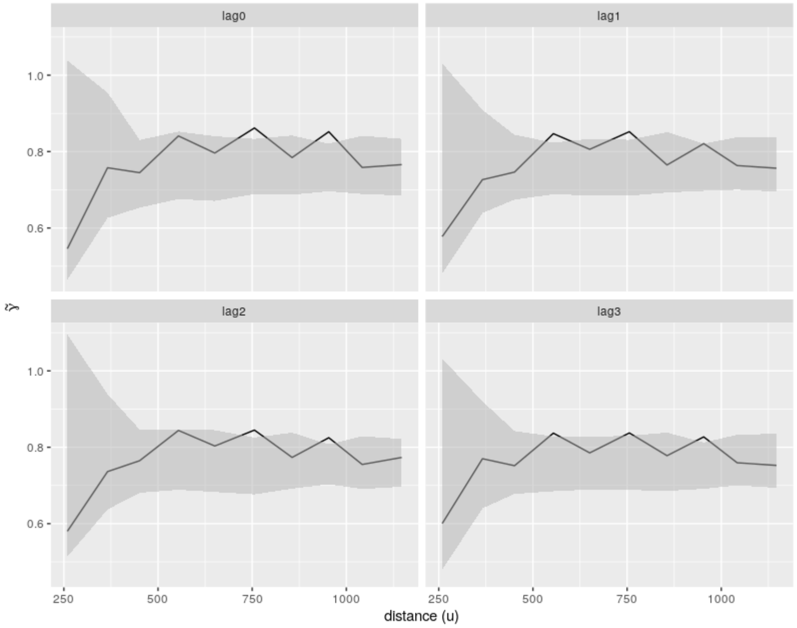

Figure 5: Spatio-temporal variogram of the residual from model including all the predictors.

above 12 per 1,000 was more important, levels of sensitivity higher than 72% could also be achieved at the expense of a specificity lower than 70%. Guidance on the choice of classifications of LSOAs with higher sensitivity should then also take into account costs of interventions, so as to identify the highest acceptable number of false alarms.

Our predictive model uses a combination of weather, environmental and socio-economic variables as predictors. Because of the high collinearity of the variables within each of the three domains and in order to make the model more parsimonious, we proposed to select a single variable to represent the risk factors domains. Grouping variables into domains of COPD risk can be particularly useful to enhance the explanatory power of the model as well as simplify the variables selection process. In our application, due to the strong collinearity among risk factors within the same domains, the differences between models considered, in the terms of predictive accuracy, were minimal.

An important finding of our analysis was the lack of evidence of residual

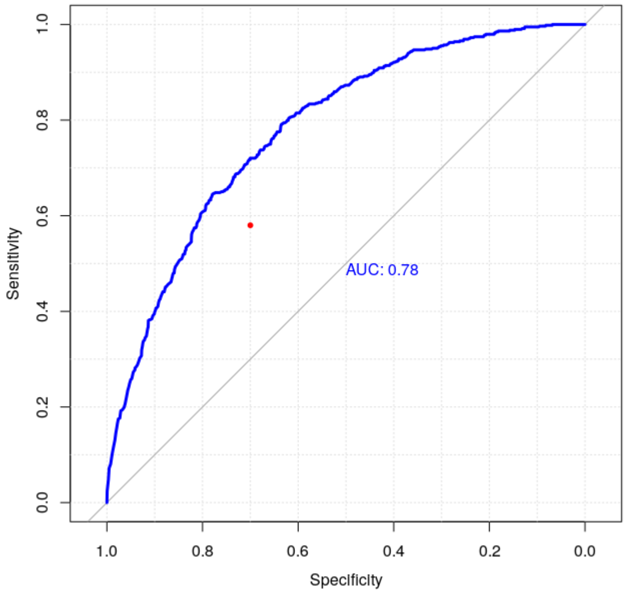

Figure 6: The receiver's characteristics curve (ROC) based on the use of exceedance probabilities (see equation (2)) for a 12 per 1,000 threshold. The red dot indicates the sensitivity and specificity for the approach based on the exceedance of the predictive mean (see equation (3)) for the admissions incidence.

spatio-temporal correlation in the reported COPD counts, after accounting for the effects of the aforementioned risk factors. If spatio-temporal correlation had been detected, our strategy would have been to model Z_{it} as a stochastic process whose sptatio-temporal correlation structure is derived from spatially continuous Gaussian process; for more explanation on the rationale and technical aspects of this approach, we refer the reader to Johnson et al. (2019).

The most important predictor in our model for COPD admissions was income deprivation. This is consistent with other studies that have reported similar findings (Calderón-Larrañaga et al., 2011; McAllister et al., 2013) and suggests that taking account of heterogeneities in socio-economic status across LSOAs is key to develop more reliable statistical models for COPD admissions.

Temporal misalignment of the predictors may be one of the main causes af-

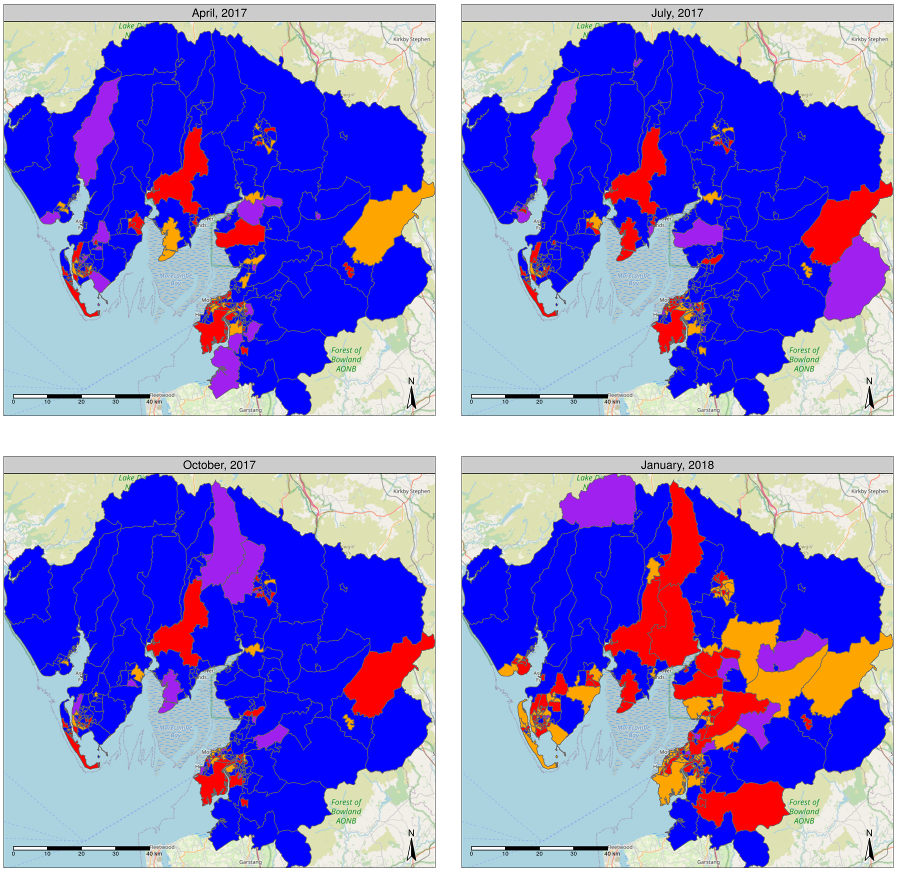

Figure 7: Maps showing Lower Super Output Areas (LOSAs) that are correctly and incorrectly classified for a COPD emergency. The interpretation of the colour coding scheme is as follows: blue indicates LSOAs that are correctly predicted to be below the threshold of 12 per 1,000 incidence rate; orange indicates an LSOA that is correctly predicted to be above the threshold; purple indicates an LSOA that is incorrectly predicted to be below the threshold; finally, red indicates an LSOA that is incorrectly predicted to be above the threshold.

fecting the predictive power of our model. Furthermore, the pollution indicators are only measured at few monitoring stations and an interpolation technique is then used to obtain estimates of pollutant concentrations across the whole of the UK. This can then further weaken the predictive performance of the best fitting model.

One of the main limitations of our study is inability to account for other important risk factors that are not available for the study region and times considered. One important missing variable in our analysis is the smoking rate, a key driver of COPD admissions, or, as an alternative proxy, lung cancer rates at LSOAs level could have also been used. Other important variables include influenza rates and the fraction of the population employed in mining and agriculture.

The developed warning system will be used to support the National Health Service Morecambe Bay Clinical Commission Group and policymakers both in the development of targeted interventions and in the resource allocation for the healthcare system, so as to reduce the need for hospital care and unplanned COPD emergency admissions. Future research will focus on improving the current modelling framework through the inclusion of risk factors that are not captured in this study.

References

Anselin, L. (1995). Local indicators of spatial association—lisa. *Geographical analysis*, *27*, 93–115.

Bahadori, K., & FitzGerald, J. M. (2007). Risk factors of hospitalization and readmission of patients with copd exacerbation–systematic review. *International journal of chronic obstructive pulmonary disease*, *2*, 241.

Bélanger, M., Couillard, S., Courteau, J., Larivée, P., Poder, T. G., Carrier, N., Girard, K., Vézina, F.-A., & Vanasse, A. (2018). Eosinophil counts in first copd hospitalizations: a comparison of health service utilization. *International Journal of Chronic Obstructive Pulmonary Disease*, *13*, 3045.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, *43*, 1–20.

Billings, J., Dixon, J., Mijanovich, T., & Wennberg, D. (2006). Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *Bmj*, *333*, 327.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, *88*, 9–25.

Calderón-Larrañaga, A., Carney, L., Soljak, M., Bottle, A., Partridge, M., Bell, D., Abi-Aad, G., Aylin, P., & Majeed, A. (2011). Association of population and primary healthcare factors with hospital admission rates for chronic obstructive pulmonary disease in england: national cross-sectional study. *Thorax*, *66*, 191–196.

Chan, F. W., Wong, F. Y., Yam, C. H., Cheung, W.-l., Wong, E. L., Leung, M. C., Goggins, W. B., & Yeoh, E.-k. (2011). Risk factors of hospitalization and readmission of patients with copd in hong kong population: analysis of hospital admission records. *BMC health services research*, *11*, 186.

Chan, T.-C., Chiang, P.-H., Su, M.-D., Wang, H.-W., & Liu, M. S.-y. (2014). Geographic disparity in chronic obstructive pulmonary disease (copd) mortality rates among the taiwan population. *PloS one*, *9*, e98170.

Donaldson, G., Seemungal, T., Jeffries, D., & Wedzicha, J. (1999). Effect of temperature on lung function and symptoms in chronic obstructive pulmonary disease. *European respiratory journal*, *13*, 844–849.

Hasegawa, W., Yamauchi, Y., Yasunaga, H., Sunohara, M., Jo, T., Matsui, H., Fushimi, K., Takami, K., & Nagase, T. (2014). Factors affecting mortality following emergency admission for chronic obstructive pulmonary disease. *BMC pulmonary medicine*, *14*, 151.

Hemming, D., Colman, A., James, P., Kaye, N., Marno, P., McNeall, D., McCarthy, R., Laing-Morton, T., Palin, E., Sachon, P. et al. (2009). Framework for copd forecasting in the uk using weather and climate change predictions.

In *IOP Conference Series: Earth and Environmental Science* (p. 142021). volume 6.

Holt, J. B., Zhang, X., Presley-Cantrell, L., & Croft, J. B. (2011). Geographic disparities in chronic obstructive pulmonary disease (copd) hospitalization among medicare beneficiaries in the united states. *International journal of chronic obstructive pulmonary disease*, *6*, 321.

Johnson, O., Diggle, P., & Giorgi, E. (2019). A spatially discrete approximation to log-gaussian cox processes for modelling aggregated disease count data. *Statistics in Medicine*, *38*, 4871–4887.

Kauhl, B., Maier, W., Schweikart, J., Keste, A., & Moskwyn, M. (2018). Who is where at risk for chronic obstructive pulmonary disease? a spatial epidemiological analysis of health insurance claims for copd in northeastern germany. *PloS one*, *13*, e0190865.

Mathers, C. D., & Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS medicine*, *3*, e442.

McAllister, D. A., Morling, J. R., Fischbacher, C. M., MacNee, W., & Wild, S. H. (2013). Socioeconomic deprivation increases the effect of winter on admissions to hospital with copd: retrospective analysis of 10 years of national hospitalisation data. *Primary Care Respiratory Journal*, *22*, 296.

Niyonsenga, T., Coffee, N., Del Fante, P., Høj, S., & Daniel, M. (2018). Practical utility of general practice data capture and spatial analysis for understanding copd and asthma. *BMC health services research*, *18*, 897.

Office for National Statistics (2018). *Mid-2017 estimates of the population for the UK, England and Wales, Scotland and Northern Ireland. 2018*. URL: `https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland` accessed 1 October 2019.

Osman, S., Ziegler, C., Gibson, R., Mahmood, R., & Moraros, J. (2017). The association between risk factors and chronic obstructive pulmonary disease in canada: A cross-sectional study using the 2014 canadian community health survey. *International journal of preventive medicine*, *8*.

Samp, J. C., Joo, M. J., Schumock, G. T., Calip, G. S., Pickard, A. S., & Lee, T. A. (2018). Predicting acute exacerbations in chronic obstructive pulmonary disease. *Journal of managed care & specialty pharmacy*, *24*, 265–279.

Tian, Y., Dixon, A., & Gao, H. (2012). Data briefing. *King's Fund, London*, .

Urwyler, P., Hussein, N. A., Bridevaux, P. O., Chhajed, P. N., Geiser, T., Grendelmeier, P., Zellweger, L. J., Kohler, M., Maier, S., Miedinger, D. et al. (2019). Predictive factors for exacerbation and re-exacerbation in chronic obstructive pulmonary disease: an extension of the cox model to analyze data from the swiss copd cohort. *Multidisciplinary respiratory medicine*, *14*, 7.

Wedzicha, J. A. (2004). Role of viruses in exacerbations of chronic obstructive pulmonary disease. *Proceedings of the American Thoracic Society*, *1*, 115–120.

Wei, X., Ma, Z., Yu, N., Ren, J., Jin, C., Mi, J., Shi, M., Tian, L., Gao, Y., & Guo, Y. (2018). Risk factors predict frequent hospitalization in patients with acute exacerbation of copd. *International journal of chronic obstructive pulmonary disease*, *13*, 121.

World Health Organisation (2016). *Chronic Obstructive Pulmonary Disease (COPD). World Health Organisation, fact sheet*. URL: `http://www.who.int/mediacentre/factsheets/fs315/en/` accessed February 6, 2018.

Yii, A. C., Loh, C., Tiew, P., Xu, H., Taha, A. A., Koh, J., Tan, J., Lapperre, T. S., Anzueto, A., & Tee, A. K. (2019). a clinical prediction model for hospitalized copd exacerbations based on "treatable traits". *International journal of chronic obstructive pulmonary disease*, *14*, 719.

Acknowledgement

We thank Professor Peter Diggle and Professor Mike Pearson for their useful comments and discussions on the paper. We also thank the University Hospitals of Morecambe Bay NHS Foundation Trust for providing the anatomised patient data.

Data Availability

The COPD data underlying the results of this study was provided by the University Hospitals Morecambe Bay NHS Foundation Trust. The data cannot be shared publicly. However, other socio-economic and environmental data are publicly available and can be made available upon request.

Funding

This research was funded by Connected Health Cities (CHC) through a PhD studentship to Olatunji. "Connected Health Cities is a Northern Health Science Alliance led programme funded by the Department of Health and delivered by a consortium of academic and NHS organisations across the north of England. The work uses data provided by patients and collected by the NHS as part of their care and support. The views expressed are those of the author(s) and not necessarily those of the NHSA, NHS or the Department of Health."

Appendix A. Table showing the performance of the combination of predictors

Table A.3: The table showing the performance of the combination of the predictors. Note RB is the rank of the bias, RR is the rank of the RMSE and Cum rank is the cumulative rank of the RB, RR and RC. Also, Min. temp is the minimum temperature, Grd frost is the number of days of ground frost and Rel. humi is the relative humidity.

ID	**Weather**	**Pollution**	**Deprivation**	**Bias**	**RMSE**	**RB**	**RR**	**Cum rank**
1	Min. temp	Income	PM_{10}	-0.172172	0.510393	1	8	9
2	Rel. humi	Education	PM_{10}	-0.172175	0.510394	2	9	11
3	Min. temp	Employment	PM_{10}	-0.172251	0.510388	36	5	41
4	Grd frost	Crime	NO_2	-0.172252	0.510386	39	3	42
5	Grd frost	Crime	PM_{10}	-0.172243	0.510395	33	10	43
6	Min. temp	Employment	NO_2	-0.172231	0.510438	25	18	43
7	Grd frost	Employment	SO_2	-0.172240	0.510433	28	16	44
8	Min. temp	Barriers	PM_{10}	-0.172252	0.510388	41	6	47
9	Grd frost	Employment	NO_2	-0.172234	0.510443	27	21	48
10	Grd frost	Income	NO_2	-0.172233	0.510444	26	22	48
11	Min. temp	Education	NO_2	-0.172243	0.510637	32	17	49
12	Grd frost	Environment	PM_{10}	-0.172242	0.510438	30	19	49
13	Min. temp	Barriers	SO_2	-0.172241	0.510439	29	20	49
14	Min. temp	Crime	PM_{10}	-0.172188	0.510576	13	36	49
15	Min. temp	Environment	NO_2	-0.172212	0.510524	21	29	50
16	Grd frost	Environment	SO_2	-0.172195	0.510572	16	35	51
17	Grd frost	Income	SO_2	-0.172291	0.510376	51	1	52
18	Rel. humi	Education	PM_{10}	-0.172291	0.510376	50	2	52
19	Min. temp	Crime	NO_2	-0.172223	0.510518	24	28	52
20	Rel. humi	Income	SO_2	-0.172212	0.510547	20	32	52
21	Min. temp	Crime	SO_2	-0.172210	0.510549	19	33	52
22	Grd frost	Environment	NO_2	-0.172288	0.510388	49	4	53
23	Min. temp	Employment	SO_2	-0.172222	0.510542	23	30	53
24	Min. temp	Barriers	NO_2	-0.172221	0.510544	22	31	53
25	Min. temp	Environment	PM_{10}	-0.172180	0.510633	7	46	53
26	Rel. humi	Environment	PM_{10}	-0.172195	0.510583	15	39	54
27	Rel. humi	Crime	NO_2	-0.172194	0.510585	14	40	54
28	Rel. humi	Environment	SO_2	-0.172242	0.510681	31	23	54
29	Grd frost	Crime	SO_2	-0.172288	0.510389	48	7	55
30	Grd frost	Education	SO_2	-0.172201	0.510577	18	37	55

Table A.4: Continuation of Table A.3

ID	**Weather**	**Pollution**	**Deprivation**	**Bias**	**RMSE**	**RB**	**RR**	**Cum rank**
31	Rel. humi	Crime	PM_{10}	-0.172201	0.510579	17	38	55
32	Min. temp	Environment	SO_2	-0.172180	0.510675	6	49	55
33	Min. temp	Income	NO_2	-0.172177	0.510693	4	54	58
34	Rel. humi	Crime	SO_2	-0.172188	0.510669	11	48	59
35	Min. temp	Health	SO_2	-0.172184	0.510677	8	51	59
36	Min. temp	Health	PM_{10}	-0.172179	0.510707	5	56	61
37	Rel. humi	Environment	NO_2	-0.172185	0.510688	9	53	62
38	Grd frost	Barriers	NO_2	-0.172176	0.510715	3	59	62
39	Min. temp	Income	SO_2	-0.172287	0.510436	47	17	64
40	Rel. humi	Education	SO_2	-0.172188	0.510702	12	55	67
41	Rel. humi	Health	PM_{10}	-0.172185	0.510711	10	58	68
42	Rel. humi	Employment	NO_2	-0.172283	0.510448	46	23	69
43	Rel. humi	Employment	PM_{10}	-0.172317	0.510432	58	14	72
44	Rel. humi	Employment	SO_2	-0.172320	0.510432	60	13	73
45	Grd frost	Barriers	PM_{10}	-0.172329	0.510412	63	11	74
46	Grd frost	Employment	PM_{10}	-0.172327	0.510413	62	12	74
47	Rel. humi	Health	SO_2	-0.172322	0.510433	61	15	76
48	Grd frost	Health	SO_2	-0.172309	0.510481	55	26	81
49	Grd frost	Barriers	SO_2	-0.172315	0.510480	57	25	82
50	Grd frost	Income	PM_{10}	-0.172319	0.510470	59	24	83
51	Grd frost	Health	PM_{10}	-0.172311	0.510484	56	27	83
52	Min. temp	Education	PM_{10}	-0.172295	0.510564	54	34	88
53	Rel. humi	Income	$NO_2$2	-0.172271	0.510618	45	43	88
54	Grd frost	Education	NO_2	-0.172264	0.510619	44	44	88
55	Min. temp	Education	SO_2	-0.172264	0.510620	43	45	88
56	Rel. humi	Barriers	NO_2	-0.172252	0.510677	38	50	88
57	Rel. humi	Barriers	SO_2	-0.172248	0.510709	34	57	91
58	Min. temp	Health	NO_2	-0.172293	0.510586	52	41	93
59	Grd frostfrost	Education	PM_{10}10	-0.172295	0.510590	53	42	95
60	Rel. humi	Barriers	PM_{10}	-0.172248	0.510717	35	60	95
61	Rel. humi	Education	NO_2	-0.172251	0.510743	37	62	99
62	Rel. humi	Health	NO_2	-0.172254	0.510730	42	61	103
63	Grd frost	Health	NO_2	-0.172252	0.510751	40	63	103