Spatial Data Analysis for Disease Mapping

Baoqi Zhang

Bachelor of Mathematics with Honours
School of Mathematics
University of Edinburgh

Abstract

This report evaluates the pressing public health issue of lung cancer in Scotland, where it is the primary cause of cancer-related mortality as well as one of the most prevalent cancer forms.

In response, our study employs disease mapping as a novel approach to better understand the spatial distribution and dynamics of lung cancer incidence across Scottish council areas. Moreover, our analysis extends to examining a range of potential covariates, including smoking rates, emission levels, earnings, and employment figures in the construction and manufacturing sectors, to uncover underlying factors contributing to lung cancer incidence. Through this approach, the report sheds light on the situation of lung cancer in Scotland, while also offers valuable insights into the complex interrelation between socio-economic factors, health behaviours, and lung cancer prevalence.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Baoqi Zhang*)

We would like to extend our warm thank you to Dr.Cecilia Balocchi for being our professor, our mentor and our light at the end of the tunnel.

Contents

Chapter 1

Introduction

> "When you view a disease as a challenge rather than a curse, you have already won half the battle."
>
> *- Joseph Lister*

Lung cancer is considered one of the most common and deadliest types of cancer and is the leading cause of cancer-related death. In Scotland, the most recent data indicates that lung cancer remains a significant health issue. In 2021, the total number of new cancer cases registered was 35,379, with lung cancer being one of the most common types (Public Health Scotland, 2023[27]). The year 2020 saw 3,874 deaths due to lung cancer in Scotland, making it the leading cause of cancer death. Interestingly, while the number of deaths from cancer in Scotland has increased over the last decade, the age-adjusted cancer mortality rate has decreased by 11% (Public Health Scotland, 2021[28]). This decrease is thought to be due to improvements in healthcare and cancer treatments. However, the absolute number of deaths due to lung cancer remains high, with nearly a quarter of all cancer deaths in Scotland attributed to lung cancer.

Given the persistent challenge lung cancer poses, it becomes crucial to explore more innovative approaches to understanding its dynamics and distribution across the population. One such approach is the utilisation of disease mapping, a technique that visualises the geographic distribution of lung cancer incidence, particularly across different council areas. It is key for understanding the spread of diseases, identifying areas at higher risk, and guiding public health interventions and resource allocation (Koch, 2005[13]). Therefore, in this report, we dedicate to obtain reliable statistical estimates of local lung cancer patterns based on the number of cases observed for each of the Scottish council areas. Additionally, analysis of potentially relevant background information such as the information and data on covariates, including smoking rate, air pollution, earnings, employment within construction and manufacturing industry, will be implemented.

This report will conduct a comparative analysis between non-spatial and spatial methodologies to ascertain their relative efficacy. Thus in the following sections, we utilise both logical regression models and the Conditional Autoregressive (CAR) models for our spatial data analysis in disease mapping. [33]

Chapter 2

Confounding

2.1 Geographical datasets

Our objective is to know the extent to which geographic factors influence the incidence and prevalence of lung cancer. To facilitate this, we have sourced a detailed shapefile delineating Scotland by council areas, an important first step in visualizing and analyzing the spatial distribution of the lung cancer count. Shapefile is a fundamental geospatial data format that encapsulates the geometric coordinates and attribute information of geographic features. This digital representation, which classifies Scotland's landscapes into points, lines, and polygons corresponding to council areas, provides the structural underpinning for our analysis[6]. Subsequent visualisation of this shapefile in Figure 2.1 facilitates hierarchical examination, demarcating each council area to enable a comprehensive exploration of spatial patterns. This study aimed to analyse the interaction between lung cancer incidence and site-specific factors, such as environmental conditions, socio-economic diversity, and lifestyle differences inherent in each council area.

2.2 Lung Cancer Incidence

Our study utilises lung cancer incidence data from 2008 to 2017, outlining cases in the Scottish council area alongside corresponding population data. This methodology enables the calculation of lung cancer incidence rates, offering a detailed analysis of the regional impact of the disease, adjusted for population size. This ratio provides a more accurate indication of lung cancer incidence, overcoming limitations inherent in raw count numbers. For instance, despite having a larger size and above-average case numbers, Highlands has significantly lower incidence rates due to its sparse population. This illustrates the importance of ratio-based analysis in comprehending the true scale of lung cancer's impact on different areas.

Our findings shown later in the report will not be affected by the COVID-19 pandemic outbreak in 2019 as our dataset only includes data from at least two years before the outbreak. Therefore, the broad socioeconomic and health impacts of COVID-19, which introduced new variables after 2019, does not affect our lung cancer analyses. This ensures that our findings are not influenced by the pandemic's transformative.

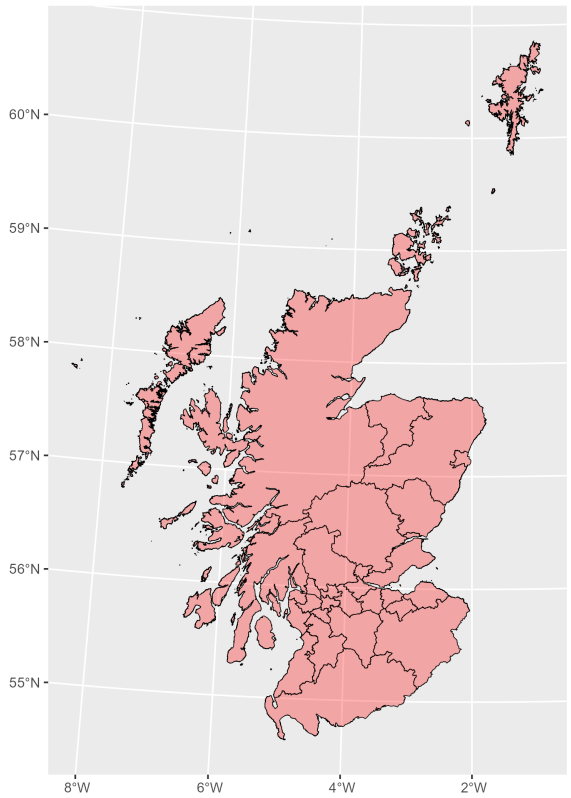


Figure 2.1: Mapping for Scotland council areas

2.3 Covariates Datasets

When analyzing the relevance of various covariates for lung cancer, we consider direct and indirect associations observed in scientific research and epidemiological studies. The reason that each covariate is selected is explained below:

Smoking is the most important and the most common etiological factor of lung cancer. Smoking can cause cell mutation, leading to the development of lung cancer. It is responsible for the vast majority of lung cancer cases and deaths. [23]. In the later analysis, the smoking rate is utilized as one of the covariates in the model, ensuring the consistency with the lung cancer incidence as discussed above, adopting a rate-based approach for data standardised.

While **Air Quality, level of CO2** is not a direct carcinogen, high level of it often indicates poor air quality. Due to a growing body of research indicating that outdoor air pollution can lead to a range of adverse health consequences, including an increased risk of lung cancer, for example, the American Cancer Society (ACS) study continues to report an increase in lung cancer cases related to air pollution[3]. we focused our subsequent data analysis on emissions per square kilometre across Scotland's council areas. This study enabled us to investigate the potential influence of environmental factors, specifically air quality, on the lung cancer incidence in these regions.

Construction industry can be a source of exposureing to hazardous substances such as asbestos, silica and diesel engine exhaust, all considered risk factors for lung cancer.

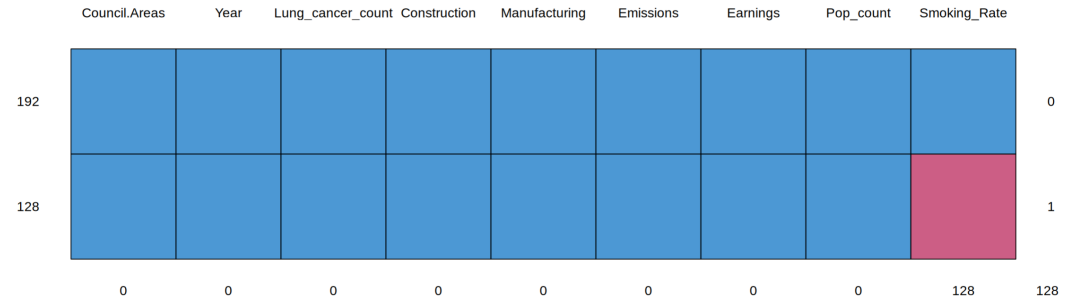
Inhaling of these carcinogens by construction workers may increase their risk of getting lung cancer. Also, the Construction Ratio in each council area in Scotland will be used in the model as for satisfying the data standardization.

Manufacturing industry, to which workers can be exposed to carcinogens such as asbestos, heavy metals, and chemical solvents. Being exposured to these substances in a manufacturing environment may increase the risk of lung cancer. Same as above, the following model will use the Manufacturing Ratio as one of the covariates for standardizing the data.

2.4 Imputation Methods for Covariates

2.4.1 Data Structure and Type

The **univariate data pattern** includes the case where there are more than 2 variables but only one variable is not completely observed [5].



Since the smoking data set is the only dataset that has missing values due to the limited periods, which is less than the standard period from 2008-2017. Hence, in this case, our data pattern is univariate.

Data are said to be Missing Completely At Random (**MCAR**)[19] if the probability of data being missing is independent of the observed and unobserved data. This can be formally represented as:

$$f(\mathbf{r} \mid \mathbf{y}, \psi) = f(\mathbf{r} \mid \psi), \quad \forall \mathbf{y}, \psi,$$

where y denotes the complete data (both observed and missing), r is a missing data indicator matrix (with entries of 1 if the data is missing and 0 otherwise), and ψ represents parameters governing the missing data mechanism. Under the MCAR assumption, the occurrence of missing data is completely unrelated to the data itself, whether observed or unobserved, and solely depends on some external parameters ψ, reflecting the overall probability of data being missing. This means that the reasons for the missing data are entirely random and unrelated to the data values.

From our original data source, the Scottish government does not display the data from 2008-2011 since it was not collected. If the absence of data is not related to any of the values that the data would have taken, then we consider the missing data as Missing Completely at Random (MCAR). This assumes that the missingness is entirely unrelated to both the observed and unobserved data.

2.4.2 Choosing Imputation Method

Initially, our code analysis employed simple mean imputation, which was ineffective as it could not adequately handle the nuances of missing data, particularly when only one variable, namely smoking rates, had missing values. We also experimented with Arima backcasting, which essentially reverses the direction of forecasting to estimate past values. However, this method predominantly relied on data from the nearest years for imputation, failing to capture the broader data trends effectively. This led us to adopt Multivariate Imputation, a more sophisticated technique that offers a robust solution by considering the relationships between multiple variables to fill in missing values.

Multivariate Imputation by Chained Equations (**MICE** [2]) is a statistical technique that is used to impute missing values iteratively according to a series of regression models. Each missing value is imputed by its own model including the process of considering each variable with missing data multiple times to account for uncertainties and relationships within the data. The MICE algorithm can impute mixes of continuous, binary, unordered categorical and ordered categorical data. In addition, MICE operates under the assumption that the missing data are Missing At Random (MAR) or MCAR.

We have chosen the Bayesian Linear Regression method as our imputation method since our smoking rate has a linear downward trend from 2012-2017. We first evaluate whether convergence has been achieved in the imputation process by observing if there is any discernible trends or systematic patterns over successive imputation iterations. We can observe that Figure 2.2 has no trend so that it has a good quality of the imputation. Furthermore, stable mean and standard deviation suggest that the imputation model has reached a point of equilibrium. Therefore, further iterations are unlikely to significantly alter the imputed values. From the Figure 2.3, we can see that the distribution of the imputed values is reasonable and there are no noticeable biases or strange artifacts. This also suggests that the imputation has been reasonably successful.

2.4.3 Aggregate Method

We use Rubin's Rule[20] to pool the results from analyses based on multiple imputed datasets that account for missing data uncertainty. It combines point estimates by averaging them across the imputed datasets for a pooled estimate. It also calculates the total variance of this estimate by incorporating both within-imputation variance (average variance within each dataset) and between-imputation variance (variance of the estimates across datasets). This approach ensures that the final statistical inferences reflect the overall uncertainty, including both the variability observed within individual imputations and the additional uncertainty introduced by the imputation process itself.

We'll use Rubin's Rules to join results from many imputations in our data analysis. It's crucial in bringing together results from different filled-in datasets since it offers a solid estimation process that considers both the variance within the imputation and extra uncertainty from missing data. Precisely, Rubin's Rule gets the average of the parameter estimates from all filled-in datasets as well as identifies the within and between-imputation variances to find out the total variance of the combined estimate.

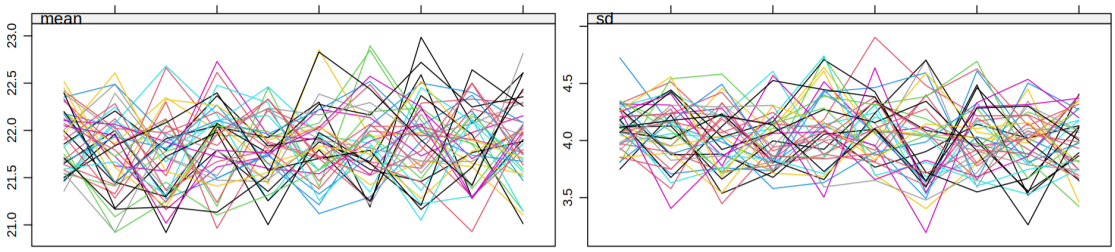


Figure 2.2: Imputation checking

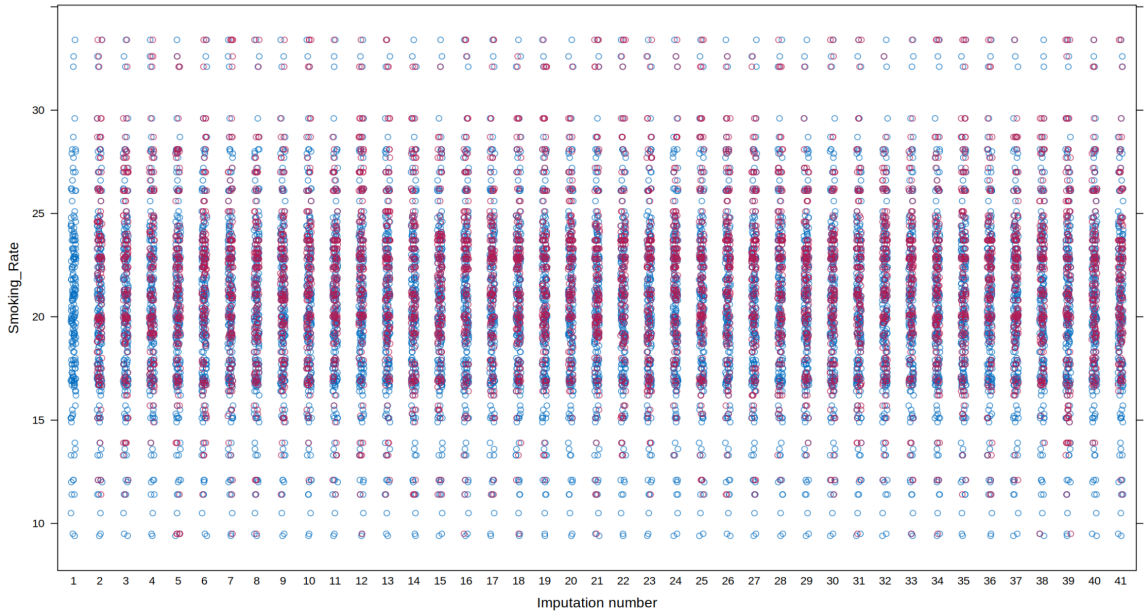


Figure 2.3: Stripplot

Chapter 3

Regression Analysis

3.1 Temporal Analysis

Regression analysis is an appropriate method for modelling an observed trend in a time series if the trend is deterministic, meaning it results from the constant, deterministic actions of a few causal forces (Jebb *et al.*, 2015[11]). After we have ensured the standardisation of all data we used in terms of time and type, the linear regression analysis is employed as the statistical tool to ascertain the existence of temporal. Analysing temporal trends in lung cancer incidence is fairly important for spatial analysis to fully understand disease dynamics. Such analyses reveal long-term patterns and short-term changes, allowing researchers to discern whether lung cancer rates are rising, falling, or staying the same. Distinguishing between long-term trends and short-term fluctuations is crucial for accurate spatial analysis, ensuring that observed spatial patterns are not attributable to transient factors. This insight is critical for measuring the effectiveness of public health initiatives, such as anti-smoking campaigns, on lung cancer prevalence.

3.1.1 Linear Regression Model

In the first place, the temporal analysis is implemented by plotting the ratios against time for the period 2008 to 2017, then accompanied by linear regression modeling. The presence of a temporal trend is determined based on the statistical significance of the p-values derived from the regression analysis.

Across the four graphs in Figure 3.1, the fluctuations observed for each line are moderate, with no sharp peaks or significant dips. Most lines keep their routes relatively limited, suggesting that while changes do occur, they are not severe. This pattern suggests that the lung cancer incidence exhibits a considerable degree of stability over the time frame described.

$$\text{Lung Cancer Incidence}_i = \beta_0 + \beta_1 \times (\text{Year} - 2012.5) + \varepsilon_i \tag{3.1}$$

The inclusion of i in the model signifies that for each council area, there is a distinct observation of lung cancer incidence ($\text{Lung Cancer Incidence}_i$), a specific year of observation (Year_i) ranged from -4.5 to 4.5, and an error term (ϵ_i). β_0 represents

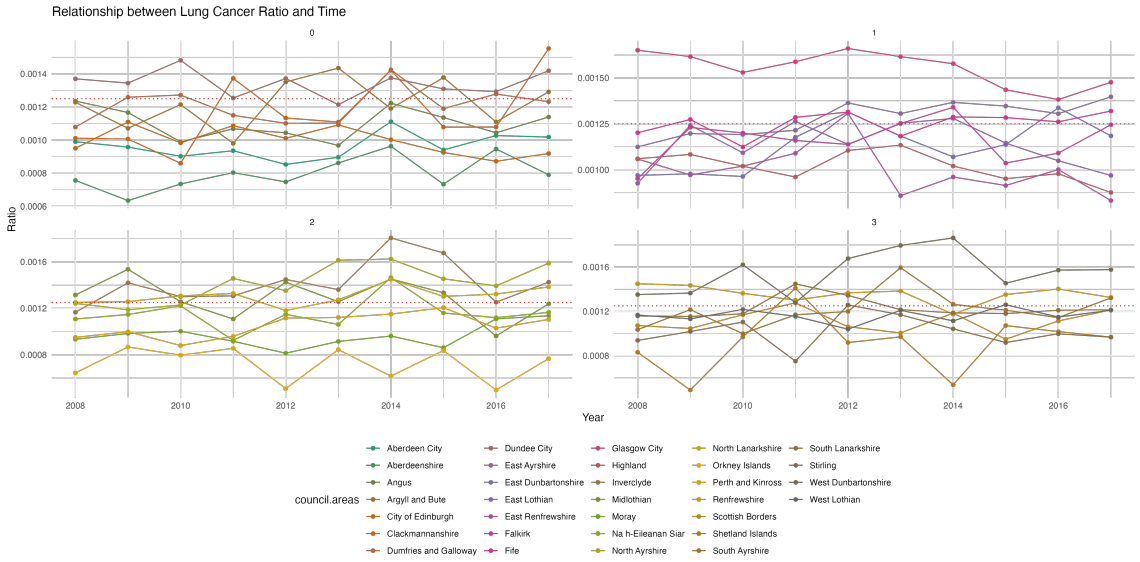


Figure 3.1: The Trend of Lung Cancer Incidence against Time

the expected value of the Lung Cancer Incidence$_i$ when the time variable Year$_i$ is at its median value of 2012.5. In other words, it is the estimated incidence rate at the midpoint of the study period, assuming all other factors influencing lung cancer rates are held constant. The parameter β_1 captures the average annual change in lung cancer incidence relative to the year 2012.5. The error term ϵ_i captures the random variation in lung cancer incidence that is not explained by the year alone, which may include other factors such as environmental, lifestyle, and genetic factors that differ from one council to another. Clustering the year variable around the median year 2012.5 (i.e., $\frac{2008+2017}{2}$) enhances the temporal analysis of lung cancer incidence by clarifying linear trends because the regression coefficients directly reflect annual changes relative to a central reference point. This approach can also mitigate multicollinearity and improve model interpretability, especially when other time-varying predictors are included.

Council	Alpha	Beta	P_Value
East Ayrshire	0.001281	0.000027	0.001096
East Dunbartonshire	0.001131	0.000030	0.038211
Glasgow City	0.001553	-0.000023	0.016590
North Ayrshire	0.001414	0.000039	0.017019
Perth and Kinross	0.001052	0.000023	0.032792

Figure 3.2: Councils with significant p-value

The linear regression analysis, focusing on lung cancer incidence trends over time and as reflected by the Beta values, reveals that the shifts in lung cancer rates across various councils are predominantly minor and gradual. From the results in Figure

3.2, out of the 32 councils analyzed, 5 have p-values that are below the significance level of 0.05, indicating statistically significant relationships between the variables in their respective regression analyses. Conversely, 27 councils have p-values above this threshold, suggesting that the evidence is not strong enough to conclude a statistically significant relationship for these councils.

3.2 Poisson Regression

3.2.1 Introducing Offset in the Model

$Y_i \sim \text{Pois}(\mu_i)$ Our regression analysis involves lung cancer incidences and five different covariates, where the data are modelled using a Poisson distribution [26] since lung cancer data is the counts which are discrete and non-negative and are often considered as rare events compared to the size of the population. Poisson regression is appropriate in this case because it accounts for the low probability of occurrence. Specifically, for each council area indexed by i, the number of lung cancer cases, μ_i, is assumed to follow a Poisson distribution: $y_i|\mu_i^* \sim \text{Pois}(\mu_i)$. Here, μ_i denotes the expected count of lung cancer incidences in different council areas. The count variable Y, representing the number of lung cancer cases, is presumed to adhere to a Poisson distribution. This distribution is uniquely determined by the parameter μ, which simultaneously signifies the distribution's mean and variance. Consequently, the probability distribution function of observing y events, with y being a non-negative integer, is articulated as:

$$P(Y_i = y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}.$$

Likelihood function L for the Poisson regression is:

$$L(\boldsymbol{\beta}; \mathbf{y}) = \prod_{i=1}^{n} P(Y_i = y_i) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$$

The natural log function serves as the link function in Poisson regression. It connects the linear predictor $\eta = \boldsymbol{\beta}^T\mathbf{x}$ to the expected value μ of the Poisson distribution. The model is expressed as:

$$\log(\mu_i) = \eta_i = \underline{\boldsymbol{\beta}}^T\underline{\mathbf{x}}_i,$$

where $\underline{\mathbf{x}}_i = (1, x_{is}, x_{im}, x_{ic}, x_{ia}, x_{ie})$ are the five covariates, i.e., Manufacturing Employment Rate, Construction Employment Rate, Emissions per hectare, Earnings per count, and Smoking Rate, meanwhile, $\underline{\boldsymbol{\beta}}_i = (\beta_0, \beta_s, \beta_m, \beta_c, \beta_a, \beta_e)$ are the coefficients for the covariates.

Since μ_i represents the count of lung cancer, and all the covariates data are in the form of a ratio. Furthermore, regions with larger populations are generally expected to have higher counts of lung cancer cases, it will be better to take the consideration the population for each council area. Without adjusting for population size, the model might misleadingly attribute differences in lung cancer counts to the other covariates when in fact they are due to differences in population size. Hence, we introduce

offset [30], which is a way to account for the population of different council areas for the dependent variable which is Lung Cancer Count in this case. By including the logarithm of the population count as an offset, we can effectively standardize the lung cancer counts by population size. This is equivalent to modelling the lung cancer rate (cases per capita) rather than the count. The offset term is included in the model without a coefficient to be estimated and it has a fixed coefficient of 1.

We then fit the model in terms of the offset and is expressed as:

$$\log(\mu_i) = \underline{\beta}^{\mathsf{T}} \underline{X}_i + \log(\text{Pop Count}_i),$$

	term	estimate	std.error	statistic	df	p.value
1	(Intercept)	-6.563278973	0.146763684	-44.720048	44.42545	1.387980e-38
2	Manufacturing	-0.009979372	0.004072174	-2.450625	157.84118	1.535260e-02
3	Construction	-0.019924923	0.006122886	-3.254172	209.46516	1.325901e-03
4	Emissions	0.009064895	0.001457523	6.219386	47.25023	1.227565e-07
5	Earnings	-0.016673112	0.003626802	-4.597194	56.89321	2.430003e-05
6	Smoking_Rate	0.013784550	0.003268649	4.217202	31.61762	1.930672e-04

Figure 3.3: Pooled Result

This Poisson regression model output indicates that increases in the Manufacturing and Construction rates are associated with decreases in lung cancer incidence, as evidenced by their negative coefficients. These relationships may potentially suggest that the physical activity associated with these industries could correlate with better overall health, which might inversely relate to lung cancer incidence. However, it is important to note that this model cannot establish causation, and the observed associations could be influenced by other factors not accounted for in the model, such as workplace safety regulations, access to healthcare, or unmeasured lifestyle variables.

The negative association between Earnings and lung cancer rates, could be reflective of socioeconomic factors where higher income levels might correlate with access to healthier lifestyles and better healthcare. This could potentially translate into a reduced risk of lung cancer. On the other hand, the pooled results show positive coefficients for both smoking rates and emissions. This aligns well with the existing literature [35] which considers smoking and air pollution as a major risk factor for lung cancer.

3.2.2 Sensitivity Analysis

The robustness of the Poisson regression model findings was evaluated through a sensitivity analysis, which aimed to ascertain the impact of imputation assumptions on the model outcomes.

Varying imputation parameters

To assess the stability of our results in the multiple imputation procedure, we varied key imputation parameters within the `mice` package, this involved altering the number of imputations. We observe from Figure 3.4 the effects on the coefficients and standard

errors for our predictors within the Poisson regression model. While there are minor fluctuations in the estimates, the direction and significance of the relationships between the covariates and lung cancer rates remained consistent, suggesting that our results are robust to changes in the imputation methodology.

```
> summary(pooled_results)
           term     estimate   std.error  statistic        df      p.value
1   (Intercept) -6.563278973 0.146763684 -44.720048  44.42545 1.387980e-38
2 Manufacturing -0.009979372 0.004072174  -2.450625 157.84118 1.535260e-02
3  Construction -0.019924923 0.006122886  -3.254172 209.46516 1.325901e-03
4     Emissions  0.009064895 0.001457523   6.219386  47.25023 1.227565e-07
5      Earnings -0.016673112 0.003626802  -4.597194  56.89321 2.430003e-05
6  Smoking_Rate  0.013784550 0.003268649   4.217202  31.61762 1.930672e-04
> summary(pooled_results_varied)
           term     estimate   std.error  statistic        df      p.value
1   (Intercept) -6.586980240 0.152010434 -43.332422  27.09468 1.455269e-26
2 Manufacturing -0.010290539 0.003937869  -2.613226 149.06474 9.888528e-03
3  Construction -0.019279670 0.006059757  -3.181591 190.54293 1.710299e-03
4     Emissions  0.008927486 0.001436957   6.212771  31.21232 6.517499e-07
5      Earnings -0.016066923 0.003726753  -4.311239  34.87766 1.264316e-04
6  Smoking_Rate  0.014152462 0.003344006   4.232188  20.10765 4.046324e-04
```

Figure 3.4: Comparison of different imputation time summary

Bootstrap Analysis

We also conduct the bootstrap analysis [10] to estimate the distribution of the model's coefficients as well as the variability in the estimates. We resampled the complete datasets with replacement, creating 1000 bootstrap samples, and refitted the Poisson regression model to each sample. The resulting distributions of the estimates for each covariate shows the extent of stability for our model. The 95% confidence intervals calculated through bootstrapping from Figure 3.5 are consistently overlapping with the estimates generated by the aggregated model.

```
                      2.5%        97.5%
Intercept     -6.895180570 -6.537672821
Manufacturing -0.016894772 -0.009534456
Construction  -0.042688920 -0.035521869
Emissions      0.005842916  0.009537830
Earnings      -0.011202950 -0.003749813
Smoking        0.009348967  0.018328946
```

Figure 3.5: 95% Confidence Interval of Bootstrap Samples

We also discover that the standard deviations of the coefficients from the multiple imputations in Figure 3.6 are typically higher than the standard errors from the pooled model for most variables. This suggests that the multiple imputation process is capturing additional uncertainty, which is not fully reflected in the pooled standard errors. This difference is an essential aspect of the multiple imputation process, which reflects the uncertainty due to missing data that is not present when analyzing a single dataset.

Since there is no direct comparative metric between Generalized Linear Models and Conditional Autoregressive Bayesian models, we then introduce the Widely Applicable

```
> print(overall_coef_sd)
[1] 0.169596434 0.007528341 0.011265115 0.001956224 0.004804426 0.003091225
```

Figure 3.6: Standard deviation for each term

Information Criterion (WAIC [4]) as an alternative to the Akaike Information Criterion (AIC) but within a Bayesian framework. This allows us for a coherent comparison between these two kinds of models, and WAIC is defined as:

$$\begin{aligned}
\text{WAIC}_2 &= 2p_{\text{WAIC2}} - 2\text{LPPD} \\
&= 2\sum_{i=1}^{N} \text{Var}_{\text{post}} \log p(y_i|\theta_i) - 2\prod_{i=1}^{N} \mathbb{E}_{\text{post}}(y_i) \\
&= 2\sum_{i=1}^{N} \text{Var}\left[\log p(y_i|\theta_i)\right] y_i - 2\log\prod_{i=1}^{N} \mathbb{E}_{\theta}\left[p(y_i|\theta_i)\right] y_i \\
&= 2\sum_{i=1}^{N} \frac{1}{M-1}\sum_{m=1}^{M}\left[\log p\left(y_i|\theta_i^{(m)}\right) - \frac{1}{M}\sum_{m=1}^{M}\log p\left(y_i|\theta_i^{(m)}\right)\right] \\
&- 2\sum_{i=1}^{N}\log\frac{1}{M}\sum_{m=1}^{M} p\left(y_i|\theta_i^{(m)}\right) \\
&= 2\sum_{i=1}^{N}\left\{\text{var}_{m=1,\ldots,M}\left[\log L_i^{(m)}\right] - \log\left(\frac{1}{M}\sum_{m=1}^{M} L_i^{(m)}\right)\right\},
\end{aligned} \tag{3.2}$$

where

$$L_i^{(m)} = p\left(y_i|\theta_i^{(m)}\right)$$

is the likelihood for the i^{th} observation given the m^{th} parameter estimates, the log pointwise predictive density (LPPD), is defined as:

$$\begin{aligned}
\text{LPPD} &= \log\prod_{i=1}^{N} p_{\text{post}}(y_i) \\
&\approx \sum_{i=1}^{N}\log\left(\frac{1}{M}\sum_{m=1}^{M} p(y_i|\theta_i^{(m)})\right),
\end{aligned}$$

and

$$p_{\text{WAIC2}} = \sum_{i=1}^{N} \text{Var}_{\text{post}}\left[\log p(y_i|\theta_i)\right].$$

Chapter 4

Spatial Data Analysis

4.1 Spatial Neighborhood Matrices

Different from the logistic regression analysis of the previous chapter, we discuss the spatial data analysis in this section. This method focuses on the lung cancer incidence spatial model in Scottish council areas. Moreover, this model is worthy of study as it can reveal the geographical relationships that may be overlooked by the previous regression model, and can provide a more comprehensive understanding of the disease patterns.

Our study focuses on area or lattice data, which implies the result of the aggregation within a defined area. In order to identify spatial relationships between different council areas, we establish neighbourhood criteria. Then we create spatial weights by assigning weights to these relationships. This stage is to verify the existence of spatial patterns in the model residuals. Some subjective evaluations should be avoided, since residual plots may sometimes suggest randomness or the existence of clusters [21]. In this chapter, we then introduce and apply rigorous methods to evaluate spatial patterns, which can provide a robust spatial dependence analysis.

Then in our study, we first use the Scotland shapefile to describe 32 different council areas, and to identify their corresponding neighbourhood areas. This is the foundational step to construct a neighbourhood matrix, which is used to assess the potential spatial autocorrelation for the disease data.

4.1.1 Neighbors based on contiguity

The definition of spatial neighbours is useful for studying the regional data and to find out whether neighbouring regions have similar or dissimilar values. Spatial neighbours can be defined in many ways, depending on the research interest and the specific setting.

Continuity-based neighbours are defined as regions that share a common border with a given region. The following Figure 4.1 shows two types of adjacency neighbours. Neighbours can be of type Queen if a single shared boundary point satisfies the proximity condition, or of type Rook if multiple shared points are required to satisfy the proximity condition.

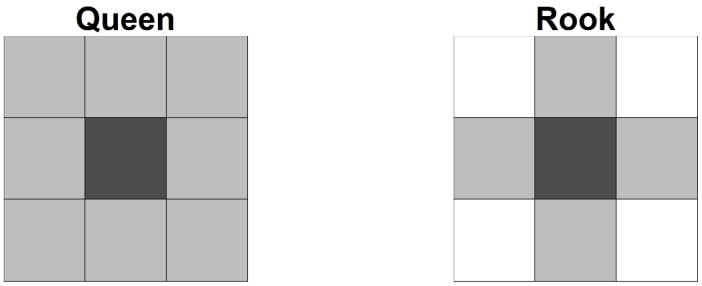


Figure 4.1: Two types of adjacency neighbours

In our analyses, we make use of the `spdep` package in R, which is used to explore spatial dependencies in data. One of the key functions in this package, `poly2nb`, specialises in illustrating spatial connections between regions by creating adjacency matrices. By default, the function constructs a weight matrix based on Rook adjacencies. The output of this function is shown in the Figure 4.2 and provides important insights into the spatial structure of the dataset. Remarkably, the output on the left reveals the fact that three areas - Nah-Eileanan Siar (20), Orkney Islands (23) and Shetland Islands (27) - lacked connections with other areas. This result reflects geographical reality: all of these islands are relatively distant from mainland of Scotland, explaining their isolation in terms of their spatial neighbours.

```
Neighbour list object:
Number of regions: 32
Number of nonzero links: 128
Percentage nonzero weights: 12.5
Average number of links: 4
3 regions with no links:
20 23 27
4 disjoint connected subgraphs
Link number distribution:

0 1 2 3 4 5 6 7 8
3 1 3 4 7 7 4 2 1
1 least connected region:
1 with 1 link
1 most connected region:
24 with 8 links
```

```
Neighbour list object:
Number of regions: 32
Number of nonzero links: 140
Percentage nonzero weights: 13.67188
Average number of links: 4.375
Link number distribution:

2 3 4 5 6 7 8
5 6 6 6 6 2 1
5 least connected regions:
8 14 19 20 27 with 2 links
1 most connected region:
24 with 8 links
```

Figure 4.2: Neighbour outputs

In the field of spatial data analysis, traditional connectivity methods, such as the k-nearest-neighbour method, may not be suitable for addressing the needs of isolated areas, such as islands, that lack direct land connections. To overcome this limitation, we analyse actual ship routes and manually link the areas where the harbour corresponding to these island routes are located. This is essential for defining the connectivity of these areas.

The investigation focuses on the transport infrastructure used by tourists and residents. It was found that the established sea routes connect the three away islands with the mainland of Scotland. Specifically, ferry services were identified connecting Orkney

and Shetland, Aberdeen City and the Highlands, services directly connecting Shetland to Aberdeen City, and some services connecting Nah-Eileanan Siar to Aberdeen City and Highland docks. The connections have been carefully mapped, as demonstrated in Figure 4.3. The resulting Figure 4.4 displays the finalised neighbour plots on the map of Scotland, ensuring integration of all Scottish council areas, including those previously considered isolated, into our network of neighbours. The varying shades of blue indicate the number of neighbours for each specific area. Additionally, the new summary of neighbours output is presented on the right-hand side of Figure 4.2. This mapping provides a comprehensive foundation for our subsequent spatial modelling analysis. It ensures that each council area is adequately represented and prepared for rigorous analysis.

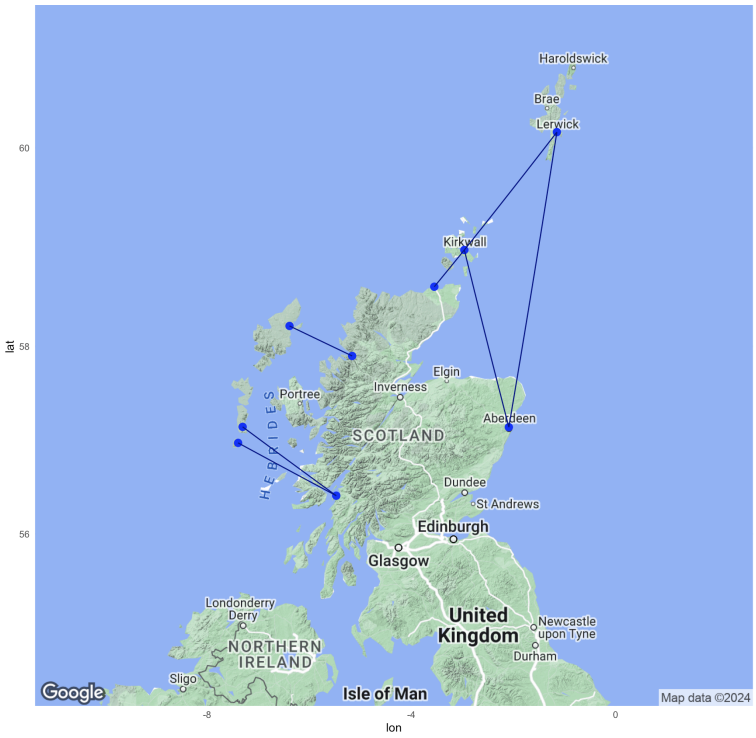


Figure 4.3: Sea routes for three isolated areas

4.1.2 Neighborhood matrices

Spatial weight is a fundamental concept in spatial analysis, which is used to quantify the spatial relationship between different areas. Weights are usually expressed in matrix format. Specifically, the connection between two areas labeled i and j is expressed numerically in the form of a spatial weight W matrix. The kth element in the ith row of this matrix corresponds to the weight or strength of the link between location i and its kth neighbor (i.e., location j) [21].

It is worth noting that the structure of these weights will vary depending on the specific method applied. For example, some methods may use binary weights [24], which indicates only the presence or absence of a relationship. Others might employ more subtle

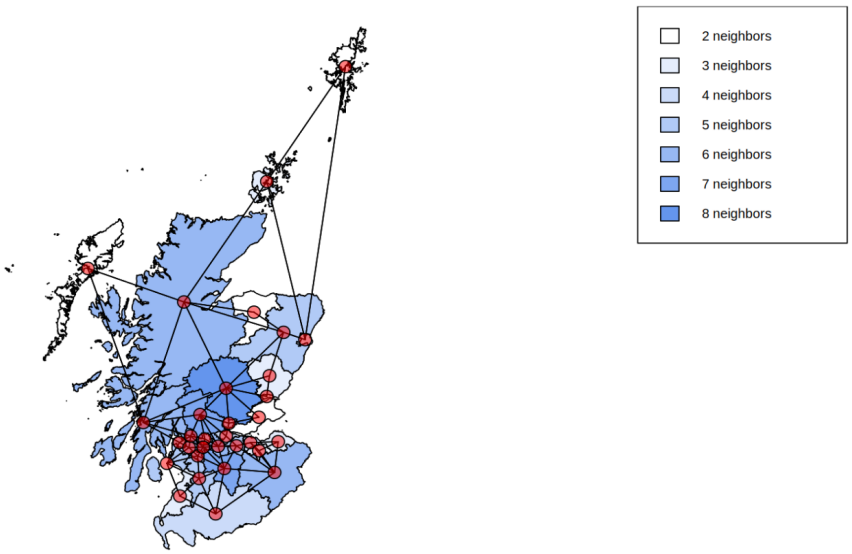


Figure 4.4: Final mapping for completed neighbours

weights, such as weights that are inversely proportional to the distance between areas, capturing the idea that closer neighbours have more influence than farther neighbours.

Since the neighbours are based on contiguity as we talked about above, we can construct a binary spatial matrix with $w_{ij} = 1$, where region i and region j share a common boundary, otherwise $w_{ij} = 0$. Customarily, w_{ii} is set to 0 for $i = 1, ..., n$. This choice of proximity measure results in a symmetric spatial matrix. Under our circumstance, the spatial matrix is a 32×32 matrix as the number of the council areas is 32. We apply the function `nb2listw` of the `spdep` package to create a spatial neighbourhood matrix containing the spatial weights corresponding to a neighbours list. The Figure 4.5 visualization of the symmetric binary weights.

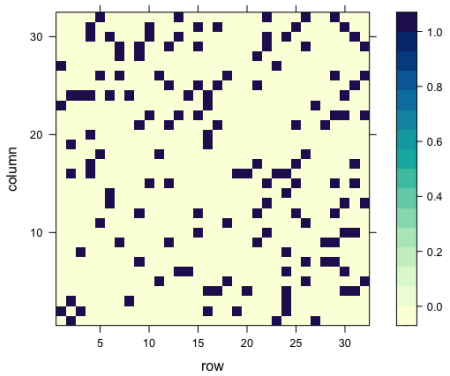


Figure 4.5: Spatial weights representation

4.2 Spatial Autocorrelation

Spatial autocorrelation refers to the degree of interdependence exhibited by a variable across a geographical space. It is fundamentally aligned with Tobler's First Law of Geography, which posits that "all things are interrelated, yet nearer objects are more strongly connected than those further apart" (Tobler, 1970 [31]). When similar values of a variable are geographically proximate, positive spatial autocorrelation is observed, manifesting as clustering. Conversely, negative spatial autocorrelation is indicated by dissimilar values being geographically proximate, leading to a dispersed pattern. Spatial autocorrelation is quantified using indices that gauge the propensity of akin observations to be geographically adjacent across a study area. Moran's I [22] and Geary's C [8] are prominent indices employed to evaluate spatial autocorrelation in area-based data. The applied procedure comes from the book [21].

4.2.1 Global Moran's I

The Global Moran's I, proposed by Moran in 1950, is an index designed to measure spatial autocorrelation by evaluating the extent of correlation between neighboring observations across a geographic region. Mathematically, it is defined as:

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \overline{Y})(Y_j - \overline{Y})}{\left(\sum_{i \neq j} w_{ij}\right) \sum_i (Y_i - \overline{Y})^2},$$

where n represents the total number of spatial units observed, Y_i denotes the lung cancer incidence in the i^{th} council area, while $\overline{Y}$ is the mean lung cancer incidence across the 32 council areas. The term w_{ij} corresponds to the spatial weight between units i and j, with the convention that $w_{ii} = 0$ and $i, j = 1, \ldots, n$, thereby ensuring that a unit does not have a spatial relationship with itself. The selection of spatial weights w_{ij} is informed by the context of the study and the characteristic spatial relationships of the variable under consideration.

The Moran's I statistic [24] serves as a tool to evaluate spatial autocorrelation, essentially measuring the similarity between a region and its neighboring regions and aggregating these comparisons. Assuming the null hypothesis, which suggests an absence of spatial autocorrelation, the observations Y_i are considered to be independently and identically distributed. Under this assumption, I follows an asymptotic normal distribution, characterized by specific mean and variance values, i.e.

$$E[I] = \frac{-1}{n-1},$$

and

$$\text{Var}[I] = \frac{n^2(n-1)S_1 - n(n-1)S_2 - 2S_0^2}{(n+1)(n-1)^2 S_0^2},$$

where

$$S_0 = \sum_{i \neq j} w_{ij}, \quad S_1 = \frac{1}{2} \sum_{i \neq j} (w_{ij} + w_{ji})^2, \quad \text{and} \quad S_2 = \sum_k \left(\sum_j w_{kj} + \sum_i w_{ik} \right)^2 .$$

Moran's I index ranges from -1 to 1. Values significantly greater than $E[I] = -\frac{1}{(n-1)}$ suggest positive spatial autocorrelation, implying that adjacent regions are likely to exhibit similar values, indicating a pattern of clustering. Conversely, values significantly less than $E[I]$ denote negative spatial autocorrelation, where neighboring regions are likely to differ in values, indicating dispersion. Values close to $E[I]$ signify a random spatial distribution, suggesting no discernible spatial pattern.

For a sufficiently large number of regions, the distribution of Moran's I approximates a normal distribution. This allows for the evaluation of spatial patterns by comparing the z-score of I to determine if the observed pattern significantly deviates from randomness. We compare the z-score

$$z = \frac{I - E[I]}{\sqrt{\text{Var}[I]}} \tag{4.1}$$

with the standard normal distribution. We reject the hypothesis that the lung cancer incidence are independent when we observe that the Moran's I statistic falls within the extreme tails of the distribution. This set of hypotheses tests for any spatial autocorrelation (positive or negative) without predetermining the direction, aligning with the concept of a two-tailed test. For testing the spatial autocorrelation, we firstly Stating the null hypothesis as well as the alternative one, i.e.

$$H_0 : I = E[I] \quad \text{(indicating no spatial autocorrelation)},$$

$$H_1 : I \neq E[I] \quad \text{(presenting spatial autocorrelation)}.$$

We then obtain the z-score value from 4.1 and set the significance level, where we normally choose 0.05 for this. The p-value is obtained when we compare the z-score value with the standard normal distribution, where we reject H_0 when the p-value is smaller than the significance level, vice versa.

Moran's I Test

As our approach is to develop the public health and epidemiological studies, hence adjusting for the population size is crucial to accurately reflect the relative frequency of an event or condition in different areas, which let us implement the rates data in this case. Furthermore, we use the mean lung cancer ratio data across from 2008-2017 for a smoothing out annual fluctuations that might be due to random variation, data collection anomalies, or short-term factors. This can provide a more stable and reliable estimate of the underlying spatial pattern of lung cancer incidence or mortality across council areas. We reduce the impact of year-to-year variability, which can obscure underlying spatial trends since our focus is on identifying consistent patterns that could indicate environmental, socioeconomic, or healthcare access factors influencing lung cancer rates.

$H_0 : I \leq E[I]$ (negative spatial autocorrelation or no spatial autocorrelation),
$H_1 : I > E[I]$ (positive spatial autocorrelation).

This set of hypotheses tests specifically for positive spatial autocorrelation, indicating a one-tailed test because the direction of the autocorrelation is predefined. We then test the normality of the lung cancer incidence across each council areas to determine if we need to conduct the Monte-Carlo randomization of the moran test, i.e. the method randomly redistributes the observed values across regions and computes Moran's I for each randomized arrangement, generating a distribution of Moran's I values under randomness. From the Figure 4.6, we observe that the W statistic is quite closed to 1 which shows that the data is following a normal distribution. Hence, we decide to conduct the moran test directly.

```
        Shapiro-Wilk normality test

data:  merged_data_1$mean_ratio
W = 0.98655, p-value = 0.952
```

Figure 4.6: Normality test

The Moran's I test output to our lung cancer ratio data is demonstrated in the Figure 4.7. We can see that the p-value is far smaller than the significance level, hence, we reject H_0 and conclude that there is a positive spatial autocorrelation. This could mean that areas with high mean ratios of lung cancer are clustered together, and the same goes for areas with low mean ratios.

```
        Moran I test under randomisation

data:  shape_sp$mean_ratio
weights: lw_q_B

Moran I statistic standard deviate = 4.126, p-value = 1.846e-05
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
       0.41183957       -0.03225806        0.01158502
```

Figure 4.7: Moran's I Statistic

Moran's I scatterplot

The Moran's I scatterplot of Figure 4.8 is to visualize the spatial autocorrelation in the mean lung cancer incidence. It illustrates how each region's data points compare to their spatially lagged counterparts. For any given region, its spatial lag is determined

by computing the weighted average of the values from adjacent regions, i.e.

$$Lag(y_i) = \sum_{j=1}^{n} w_{ij} y_j,$$

where $Lag(y_i)$ is the spatial lag of variable y at location i, w_{ij} is the weight that reflects the spatial relationship or influence between location i and location j, y_j is the value of y at location j, and n is the total number of locations. This plot reveals a positive linear relationship between the observations and their spatially lagged values. Using this plot, we can also identify the data points which are closer to the fitted line have a high influence of the spatial pattern.

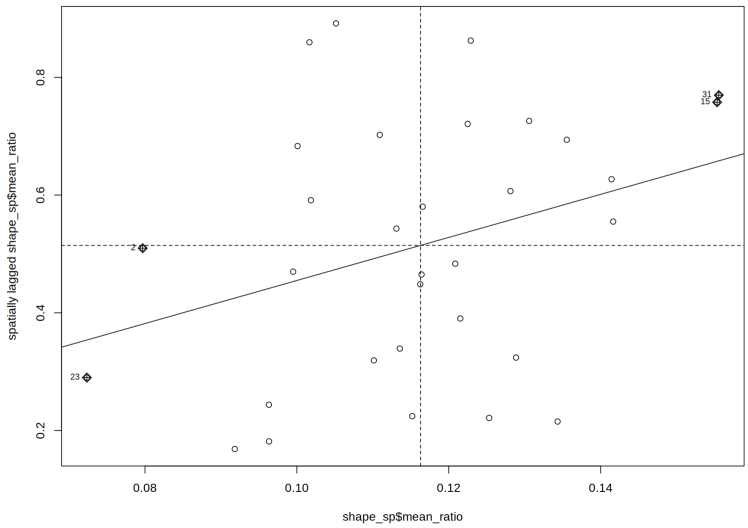


Figure 4.8: Moran's Plot

4.2.2 Local Moran's I

Local Indicators of Spatial Association (LISA) aim to identify significant spatial clustering of similar values around each observation. They possess the advantageous characteristic of their total sum across all observations being equivalent to a multiple of the comprehensive measure of spatial association. This means global spatial autocorrelation statistics can be broken down into individual and local components. Essentially, many LISA measures are localized adaptations of broader, globally recognized spatial autocorrelation indices.

The Local Moran's I of the i^{th} region is shown as:

$$I_i = \frac{n}{\sum_j (Y_j - \bar{Y})^2} (Y_i - \bar{Y}) \sum_j w_{ij} (Y_j - \bar{Y}).$$

It's important to recognize that the overall Moran's I is directly related to the cumulative sum of local Moran's I values calculated for every region:

$$I = \frac{1}{\sum_{i \neq j} w_{ij}} \sum_{i} I_i.$$

The LISA values are visualized on a map to show the locations of areas with relatively high or low local connectivity to their neighbors. A high I_i value indicates that the area is encircled by others with comparable values, positioning it within a cluster of either high, low, or moderate observations. Conversely, a low I_i value suggests the area is surrounded by areas with contrasting values, marking it as an outlier. This signifies that the observation in area i stands out from the majority or all neighboring observations. The subsequent two diagrams exemplify the local variant of the Moran's I test.

Parameter Plot

For understanding the local Moran's I for each area, we generate a map of p-values that illustrate the likelihood of surpassing the observed values under the assumption that the null hypothesis holds. These p-values can be determined irrespective of whether there is a global spatial association or not, through a simulation method employing a conditional randomization strategy. In this strategy, the observed value Y_i, representing the average lung cancer ratio in region i, is kept constant while the other values are randomly redistributed among the remaining regions.

We employ the `tmap` function to visually present the rates of lung cancer, local Moran's I, z-scores, and p-values for various council areas. Regions exhibiting a p-value below the significance threshold of 0.05 (or with z-scores exceeding qnorm(0.95) = 1.65) are identified as areas where the null hypothesis is rejected, indicating the presence of positive spatial autocorrelation.

We then interpretating the 4 plots in the Figure 4.9:

Lung Cancer Rate: This map displays the average ratio by population of lung cancer cases between 2008 and 2017, categorized across Scottish council areas. The shading indicates the ratio, with a colour gradient from light yellow (denoting a low rate of lung cancer) to dark orange (indicating a higher rate of lung cancer). Noticeably, 'Glasgow City' and 'West Dunbartonshire' have notable figures, in contrast to 'Orkney Islands,' which is marked by light yellow to denote the lowest incidence rates.

Local Moran's I: This map shows the Local Moran's I statistic for each area, which identifies local clusters of similar values (high or low). Green shades indicate positive spatial autocorrelation (clusters of similar high or low values), while orange shades suggest negative spatial autocorrelation (outliers or locations with dissimilar values compared to their neighbours).

Z-score: The Z-score map indicates areas of statistical significance in terms of spatial clustering. Red regions are those where the Z-score is above a certain threshold (1.645), suggesting that the observed spatial clustering is statistically significant and not due to random chance.

P-value: This map provides a significance test for the presence of spatial autocorrelation. Regions in red have a p-value less than 0.05, indicating that the probability of

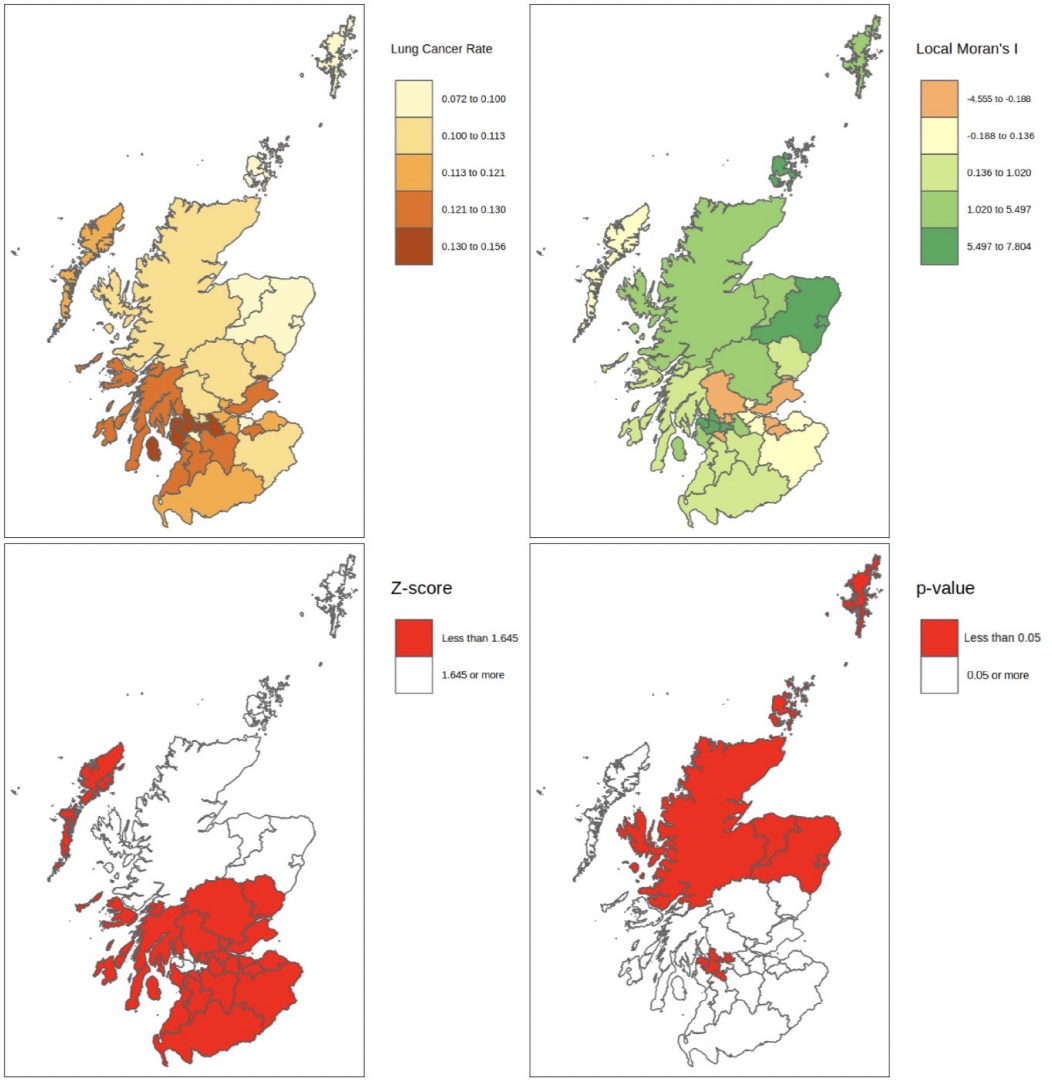


Figure 4.9: Local Moran's and Hypothesis Test Plots

observing such a spatial pattern by random chance is less than 5%, and thus the pattern is statistically significant.

Lisa Cluster Plots

Local Indicators of Spatial Association (LISA) cluster plots are a type of statistical analysis used in geographic information systems to identify areas with either high or low values of a particular variable that are spatially clustered together. In the context of analyzing the mean lung cancer ratio, a LISA cluster plot can serve several purposes: Interpretations:

- High-High areas (Red areas): These indicate clusters where locations with high lung cancer incidence are surrounded by other locations with similarly high lung cancer incidence. This implies positive spatial autocorrelation, suggesting that there are hotspots where lung cancer incidences are consistently high across

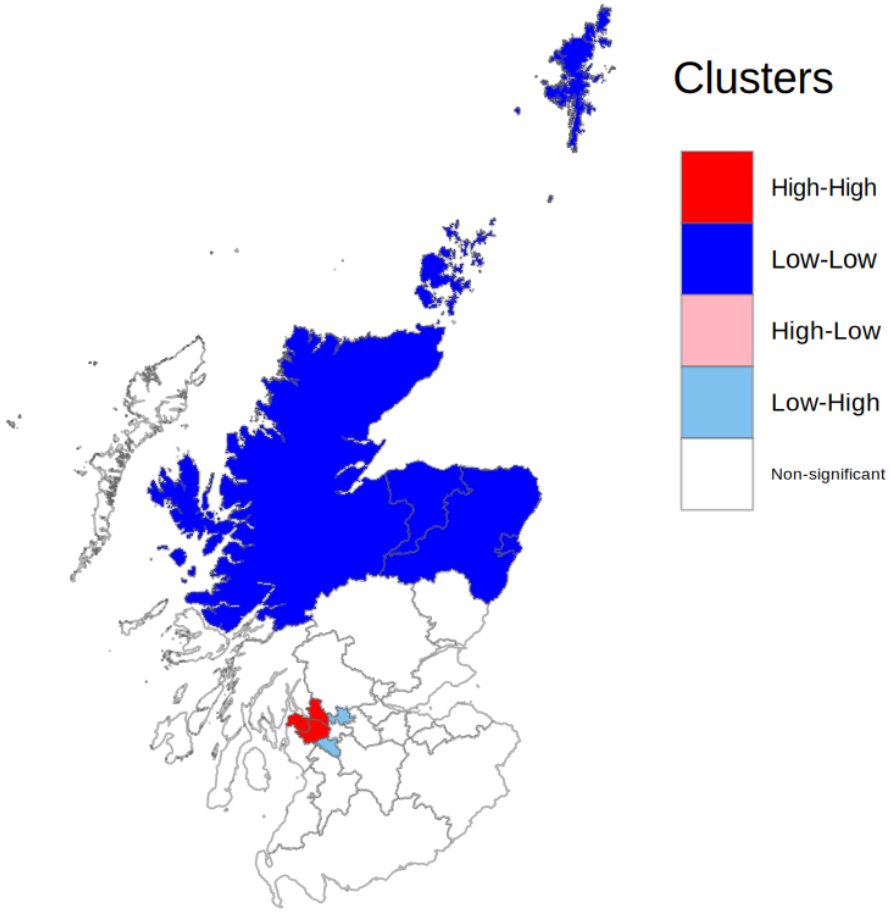


Figure 4.10: LISA Cluster Plot

neighbouring areas. The three red areas are 'West Dunbartonshire','Inverclyde' and 'Renfrewshire' respectively.

- Low-Low areas (Blue areas): These are clusters where locations with low lung cancer incidence are surrounded by other locations with low lung cancer incidence. This also reflects positive spatial autocorrelation, but in this case, it indicates cold spots where lung cancer incidences are consistently low across neighbouring areas. From the graph, we can observe that the north part of Scottish council areas generally has a low lung cancer incidence.

- High-Low areas (Light Pink areas): These locations have high lung cancer incidence but are surrounded by locations with low lung cancer incidence. They could represent unique environmental or social factors that affect lung cancer rates differently than in surrounding areas, showing negative spatial autocorrelation as spatial outliers.

- Low-High areas (Light Blue areas): These areas have low lung cancer incidence despite being surrounded by locations with high lung cancer rates. They may also be considered spatial outliers, suggesting negative spatial autocorrelation

and potentially different underlying factors influencing the lower rates in these areas ('East Dunbartonshire' and 'East Renfrewshire').

- Non-significant areas (White areas): These regions do not show any significant spatial autocorrelation for lung cancer incidence at the chosen significance level (commonly 0.05). This means that the lung cancer rates in these areas do not form a statistically significant cluster pattern and may appear randomly distributed.

When looking at such a map, public health officials might focus interventions or further investigations in the High-High clusters to understand the causes of elevated lung cancer rates and to develop targeted health programs. Conversely, the Low-Low clusters might be examined to identify protective factors or effective prevention strategies.

4.3 Bayesian Spatial Models

In exploring spatial relationships and dependencies through the perspective of neighbours and spatial weights, a fundamental understanding for mapping disease distribution is established. The focus then shifts to an underlying analytical tool that utilises the concept: **Conditional Autoregressive** (CAR) model. CAR models are hailed as key elements in the field of spatial statistics and are particularly relevant for disease mapping because their ability to accommodate the spatial autocorrelation inherent in geographic data is outstanding. By incorporating spatial structure delineated by neighbours and weights, CAR models support complex analyses in which disease incidence in one region may be influenced by disease incidence in neighbouring regions. Hence, this approach not only improves the accuracy of disease mapping but also improves the understanding of the spatial dynamics of disease spread, validating the CAR model as an important tool in the spatial data analysis toolbox for disease mapping.

The study domain $\mathbf{S}$, representing Scotland, is segregated into $\mathbf{K} = 32$ discrete non-overlapping spatial units, corresponding to the 32 council areas of Scotland. Each associated with a respective set of outcomes $\mathbf{Y} = (Y_1, \ldots, Y_K)$, and a pre-determined offsets vector $\mathbf{O} = (O_1, \ldots, O_K)$. Here, the offsets stand for a population of the 32 council areas. Spatial variation in the responses is captured by a covariates matrix $\mathbf{X} = (X_1, \ldots, X_K)$ and a spatial structural element $\psi = (\psi_1, \ldots, \psi_K)$, the latter for modelling remaining spatial autocorrelation post-covariate adjustment. Covariate vectors for spatial unit S_k are represented as $\boldsymbol{x_k} = (1, x_{k1}, \ldots, x_{kp})$, with the leading term corresponding to an intercept and p denoting the number of covariates, which in this case is 5. These include Smoking Rate, Manufacturing Employment, Construction Employment, Emissions, and Earnings. The **CARBayes**[17] package is equipped to fit several data likelihood models, such as Binomial, Gaussian, Poisson, and Zero-Inflated Poisson (ZIP). In the case of modelling lung cancer incidence and relevant covariate data, the Poisson model emerges as the most suitable fit among the options provided by the package (Lee, 2020[16]). Furthermore, to capture the temporal dynamics of lung cancer incidence more accurately, our analysis differentiates between "Single Year" and "Multiple Year" models within the CAR framework. Specifically, it is modelled using data for individual years to understand annual variations ("Single Year" approach)

and aggregate data over multiple years to identify longer-term trends ("Multiple Year" approach).

Poisson (Single Year): $Y_k \sim \text{Poisson}(\mu_k)$ and $\log(\mu_k) = \boldsymbol{x}_k^T\boldsymbol{\beta} + O_k + \psi_k$, where

$$\psi_k = \phi_k + \theta_k;$$

Poisson (Multiple Years): $Y_{kt} \sim \text{Poisson}(\mu_{kt})$ and $\log(\mu_{kt}) = \boldsymbol{x}_{kt}^T\boldsymbol{\beta} + O_{kt} + \psi_{kt}$, where

$$\psi_{kt} = \phi_{kt} + \theta_{kt};$$

and

$$\begin{aligned}(\phi_{kt})_{k=1,\ldots,n} = (\phi_{1t}, ..., \phi_{nt}) &\sim \text{CAR}(\rho, \tau^2),\\ (\theta_{kt})_{k=1,\ldots,n} = (\theta_{1t}, ..., \theta_{nt}) &\overset{\text{iid}}{\sim} \text{N}(0, \tau^2),\\ &\forall t = 1, ..., T.\end{aligned}$$

The report will incorporate two established CAR priors: the Besag-York-Mollié (BYM) model (Besag *et al.*, 1991[1]), and the alternative model Leroux developed by (Leroux *et al.*, 2000[18]). A comparative analysis of these models' performance in spatial data interpretation will also be conducted.

Typically, the spatial structure component ψ includes a set of random effects $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_K)$, derived from a conditional autoregressive model. These models generally take the form $\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{0}, \tau^2\mathbf{Q}(\mathbf{W})^{-1})$, where $\mathbf{Q}(\mathbf{W})$ is the precision matrix that may be singular (e.g., in the intrinsic model). This matrix controls the spatial autocorrelation of the random effects and is based on a non-negative symmetric $K \times K$ neighbourhood matrix $\mathbf{W}$, which is discussed above. Moreover, CAR priors are typically characterized by a series of K univariate full conditional distributions $f(\phi_k|\boldsymbol{\phi}_{-k})$ for $k = 1, \ldots, K$, where $\boldsymbol{\phi}_{-k} = (\phi_1, \ldots, \phi_{k-1}, \phi_{k+1}, \ldots, \phi_K)$.

In an effort to provide a more straightforward and intuitive understanding of the CAR modelling process, we can utilize a matrix regression framework. This approach encapsulates the model in a familiar regression format. Meanwhile, $\mathbf{Y}$ represents the lung cancer cases, and $\boldsymbol{\alpha}$ corresponds to the intercept term. The term $\mathbf{X^T}\boldsymbol{\beta}$ denotes the product of the transpose of the covariate matrix $\mathbf{X}$ and the vector $\boldsymbol{\beta}$, which contains the coefficients for the covariates. This regression structure enables the model to handle spatial and temporal dependencies systematically, translating complex multivariate relationships into a comprehensible linear algebraic form.

$$\mathbf{Y} = \boldsymbol{\alpha} + \mathbf{X^T}\boldsymbol{\beta} \tag{4.2}$$

Then first consider the single-year model, the vector $\mathbf{Y}$ is the response variable vector with dimensions 32×1, reflecting single-year lung cancer counts across 32 council areas. The vector $\boldsymbol{\alpha}$ captures baseline levels for the different council areas, which has the dimension 32×1, each entry representing a unique council area.

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{32} \end{pmatrix}, \quad \boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{32} \end{pmatrix},$$

The matrix $\mathbf{X}$ has dimensions 5×32, representing 5 covariates for each council area, and upon transposition becomes 32×5. The vector $\boldsymbol{\beta}$ contains the coefficients for the covariates and is of the dimension 5×1. The covariates are defined as follows: s for Smoke Rate, m for Manufacturing Employment, c for Construction Employment, e for Emission, and a for Earning.

$$\mathbf{X} = \begin{pmatrix} x_{1,s} & x_{2,s} & \cdots & x_{5,s} & x_{6,s} & \cdots & x_{32,s} \\ x_{1,m} & x_{2,m} & \cdots & x_{5,m} & x_{6,m} & \cdots & x_{32,m} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{1,e} & x_{2,e} & \cdots & x_{5,e} & x_{6,e} & \cdots & x_{32,e} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_s \\ \beta_m \\ \beta_c \\ \beta_a \\ \beta_e \end{pmatrix}$$

Then consider the multi-year model, the vector $\mathbf{Y}$ is an aggregation of the response variable across several years, leading to a dimension of $32 \cdot t \times 1$, where t is the number of years that we consider to take into account in the model. It reflects data from 32 council areas over $t = 10$ years, now reshaped into a column vector for analysis. The vector $\boldsymbol{\alpha}$ maintains its role as the intercept, now extended to account for each year's baseline level for all council areas. It also transforms into a $32 \cdot t \times 1$ vector, aligning with the reshaped $\mathbf{Y}$.

$$\mathbf{Y} = \begin{pmatrix} y_{1,1} \\ \vdots \\ y_{32,1} \\ y_{1,2} \\ \vdots \\ y_{32,2} \\ \\ \vdots \\ \\ y_{1,t} \\ \vdots \\ y_{32,t} \end{pmatrix}, \quad \alpha = \begin{pmatrix} \alpha_{1,1} \\ \vdots \\ \alpha_{32,1} \\ \alpha_{1,2} \\ \vdots \\ \alpha_{32,2} \\ \\ \vdots \\ \\ \alpha_{1,t} \\ \vdots \\ \alpha_{32,t} \end{pmatrix}$$

The covariate matrix $\mathbf{X}$ broadens to include data across all years, yielding a $5 \times 32 \cdot t$ matrix. Each block of 5×32 within $\mathbf{X}$ corresponds to a year's worth of covariates for all council areas, stacking horizontally for each year.

$$\mathbf{X} = \begin{pmatrix} x_{1,1,s} & \cdots & x_{32,1,s} & x_{1,2,s} & \cdots & x_{32,2,s} & \cdots & x_{1,t,s} & \cdots & x_{32,t,s} \\ x_{1,1,m} & \cdots & x_{32,1,m} & x_{1,2,m} & \cdots & x_{32,2,m} & \cdots & x_{1,t,m} & \cdots & x_{32,t,m} \\ x_{1,1,c} & \cdots & x_{32,1,c} & x_{1,2,c} & \cdots & x_{32,2,c} & \cdots & x_{1,t,c} & \cdots & x_{32,t,c} \\ x_{1,1,a} & \cdots & x_{32,1,a} & x_{1,2,a} & \cdots & x_{32,2,a} & \cdots & x_{1,t,a} & \cdots & x_{32,t,a} \\ x_{1,1,e} & \cdots & x_{32,1,e} & x_{1,2,e} & \cdots & x_{32,2,e} & \cdots & x_{1,t,e} & \cdots & x_{32,t,e} \end{pmatrix}$$

Vector $\boldsymbol{\beta}$ remains a 5×1 vector, containing the coefficients for the covariates, but now encapsulates the influence of these covariates across all years.

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_s \\ \beta_m \\ \beta_c \\ \beta_a \\ \beta_e \end{pmatrix}$$

The transpose of $\mathbf{X}$, denoted by $\mathbf{X^T}$, has dimensions $32 \cdot t \times 5$, facilitating the multiplication with the vector $\boldsymbol{\beta}$ to compute the linear predictor for the expanded model.

$$\mathbf{X^T} = \begin{pmatrix} x_{1,1,s} & x_{1,1,m} & x_{1,1,c} & x_{1,1,a} & x_{1,1,e} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{32,1,s} & x_{32,1,m} & x_{32,1,c} & x_{32,1,a} & x_{32,1,e} \\ x_{1,2,s} & x_{1,2,m} & x_{1,2,c} & x_{1,2,a} & x_{1,2,e} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{32,2,s} & x_{32,2,m} & x_{32,2,c} & x_{32,2,a} & x_{32,2,e} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1,t,s} & x_{1,t,m} & x_{1,t,c} & x_{1,t,a} & x_{1,t,e} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{32,t,s} & x_{32,t,m} & x_{32,t,c} & x_{32,t,a} & x_{32,t,e} \end{pmatrix}$$

It then appears that

$$\mathbf{Y} = \boldsymbol{\alpha} + \mathbf{X^T}\boldsymbol{\beta} = \begin{pmatrix} y_{1,1} \\ \vdots \\ y_{32,1} \\ y_{1,2} \\ \vdots \\ y_{32,2} \\ \vdots \\ y_{1,t} \\ \vdots \\ y_{32,t} \end{pmatrix} = \begin{pmatrix} \alpha_{1,1} \\ \vdots \\ \alpha_{32,1} \\ \alpha_{1,2} \\ \vdots \\ \alpha_{32,2} \\ \vdots \\ \alpha_{1,t} \\ \vdots \\ \alpha_{32,t} \end{pmatrix} + \begin{pmatrix} x_{1,1,s} & x_{1,1,m} & x_{1,1,c} & x_{1,1,a} & x_{1,1,e} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{32,1,s} & x_{32,1,m} & x_{32,1,c} & x_{32,1,a} & x_{32,1,e} \\ x_{1,2,s} & x_{1,2,m} & x_{1,2,c} & x_{1,2,a} & x_{1,2,e} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{32,2,s} & x_{32,2,m} & x_{32,2,c} & x_{32,2,a} & x_{32,2,e} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1,t,s} & x_{1,t,m} & x_{1,t,c} & x_{1,t,a} & x_{1,t,e} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{32,t,s} & x_{32,t,m} & x_{32,t,c} & x_{32,t,a} & x_{32,t,e} \end{pmatrix} \begin{pmatrix} \beta_s \\ \beta_m \\ \beta_c \\ \beta_a \\ \beta_e \end{pmatrix} \tag{4.3}$$

Moreover, in the bayesian model for lung cancer incidence, we use Markov Chain Monte

Carlo (MCMC) simulations by employing three chains to enhance the robustness of the estimate. Markov Chain Monte Carlo (MCMC) methods generate dependent samples from a target distribution, $p(\theta)$, or for Bayesian inference $p(\theta|y)$ [7], through a Markov chain. This chain has a limiting stationary distribution equal to the target distribution, implying the chain values eventually converge to be a sample from $p(\theta)$. Running 3 chains with different starting values allow us to confirm the chains are converging to the desired target distribution.

4.3.1 Choosing the Representative Imputation Data to perform CAR

Challenges with Pooling in Bayesian CAR Models:

- Complexity of Models: Bayesian CAR models, often used for spatial data analysis, involve complex dependencies and priors. Integrating multiple imputations into this framework can add considerable computational complexity and difficulty in interpreting pooled results.
- Pooling Uncertainty: Traditional pooling methods, such as Rubin's Rules, are not directly applicable to Bayesian models because Bayesian analyses inherently account for parameter uncertainty. The challenge is in combining the uncertainty from the imputation process with the posterior distributions obtained from the Bayesian analysis.

In our analysis, we sought to address missing data by employing multiple imputation via the mice package, generating 40 different complete datasets. We choose randomly from the imputation datasets for our further analysis as we can observe that the result of the imputation data only varies in a small scale and the standard deviation is extremely small from Figure 3.6.

4.3.2 CARbym Modelling

The convolution approach in the BYM CAR model integrates spatially autocorrelated and independent random effects. This methodology, as established in Besag *et al.* (1991)[1], allows for a comprehensive representation of spatial variation in the data. It is given by

$$\begin{aligned} \psi_k &= \phi_k + \theta_k \\ \phi_k|\phi_{-k}, \mathbf{W}, \tau^2 &\sim \mathcal{N}\left(\frac{\sum_{i=1}^{K} w_{ki}\phi_i}{\sum_{i=1}^{K} w_{ki}}, \frac{\tau^2}{\sum_{i=1}^{K} w_{ki}}\right) \\ \theta_k &\sim \mathcal{N}(0, \sigma^2) \\ \tau^2, \sigma^2 &\sim \text{Inverse-Gamma}(a, b) \end{aligned} \tag{4.4}$$

This formula is for the single-year BYM model. In this framework, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ represents independent random effects with a mean of zero and consistent variance.

Spatial autocorrelation is captured through random effects $\phi = (\phi_1, \ldots, \phi_K)$. The expected values of ϕ are calculated as the average of the random effects from adjacent areas, and the variance of ϕ is inversely proportional to the number of neighbouring areas. This formulation is justified as increased spatial autocorrelation of random effects leads to more information being available from neighbouring areas, thereby reducing uncertainty. In line with other variance parameters, the default prior setting for (τ^2, σ^2) is given by $a = 1, b = 0.01$. The model incorporates two random effects for each data point. However, since the data reveals only the aggregate effect, the user is provided with the sum $\psi_k = \phi_k + \theta_k$ for each point.

CARbym Single Year

	Mean	2.5%	97.5%	n.sample	% accept	n.effective	PSRF (upper 95% CI)
(Intercept)	-7.3019	-8.3136	-6.2971	1e+05	45.1	4125.9	1
Smoking_Rate	0.0266	0.0044	0.0486	1e+05	45.1	3401.1	1
Manufacturing	0.0073	-0.0301	0.0457	1e+05	45.1	4645.5	1
Construction	-0.0120	-0.0637	0.0392	1e+05	45.1	4810.0	1
Emissions	-0.0002	-0.0129	0.0129	1e+05	45.1	2870.8	1
Earnings	0.0007	-0.0237	0.0254	1e+05	45.1	5854.9	1
tau2	0.0165	0.0040	0.0424	1e+05	100.0	8438.4	1
sigma2	0.0049	0.0016	0.0124	1e+05	100.0	8984.3	1

Figure 4.11: Result from running CARbym function (2016)

The first BYM model, constructed from the 2016 dataset, allows us to investigate predictors of lung cancer incidence. The model's `intercept`, estimated at -7.3019, provides a reference point against which the effects of the covariates are measured. Notably, each covariate's effect is interpreted conditionally, acknowledging that other variables in the model are held constant. The coefficient for `Smoking Rate`, at 0.0266, implies that keeping other variables steady, an increase in the smoking rate is associated with an increase in lung cancer incidence. This finding is consistent with extensive research highlighting smoking as a primary lung cancer risk factor. It is important to interpret this coefficient in the context of the model's other variables, as the actual impact of smoking could be confounded or moderated by these factors.

For `Manufacturing Employment count`, with a coefficient of 0.0073, the interpretation suggests that, conditional on the other variables, higher employment in manufacturing is associated with a slightly increased risk of lung cancer. This could point to occupational exposures, and hence, it warrants an in-depth examination of health risks in manufacturing settings. The negative association indicated by the `Construction Employment count` coefficient, at -0.0120, is interesting. It implies that, when controlling for other factors, higher construction employment might correlate with a lower incidence of lung cancer. This could reflect a protective effect or the presence of other unmeasured variables that correlate with employment in construction and influence lung cancer incidence.

The near-zero coefficient for `Emissions`, at -0.0002, should be interpreted with caution. This coefficient suggests that assuming other variables in the model are held constant,

changes in Emissions have a negligible direct effect on lung cancer incidence. This implies that within the context of this particular model – which includes variables such as Smoking Rate, Manufacturing, and Construction Employment counts – Emissions alone do not significantly contribute to explaining the variability in lung cancer rates. However, this does not rule out the possibility that Emissions could have an indirect effect or be important in a different modelling context, especially when considering potential multicollinearity among predictors.

The `Earnings` coefficient, at 0.0007, seemingly contradicts the anticipated negative correlation between income and adverse health outcomes. This implies that within the context of the other model variables, higher earnings are associated with a slightly higher incidence of lung cancer. This counterintuitive result could reflect complex socio-economic patterns or specificities in the dataset and warrants further examination. However, a negative coefficient for `Earnings` in the multi-year model will turn out in the next section. It will align with the broader literature and the generally accepted understanding that higher socioeconomic status is often associated with better health outcomes.

Finally, the variance components — `tau2` at 0.0165 and `sigma2` at 0.0049 — emphasize the existence of spatial dependence in lung cancer incidence rates. A considerable proportion of the variation in incidence rates is due to spatially structured random effects, highlighting the non-uniform geographic distribution of lung cancer risk. This reinforces the necessity for geographically nuanced public health strategies and interventions.

For the model check of this BYM single-year model, the trace plots in the Figure A.1 of the model parameters show stable variance and no apparent trend or drift during the iterations. This behaviour supports the fact that the chain has reached equilibrium and the parameters are estimated, so it can be considered reliable. The density plot on the left side of Figure A.1 is bell-shaped and centred on the mean of the sampled values, which suggests that the posterior distribution is well-defined and the parameter distributions are clear and stable. Also, the results of the Gelman-Rubin diagnostic are shown in Figure A.2 and its Gelman-Rubin test result is 1. This means that the variance within each chain is similar to that between different chains and that the chains are effectively sampling from the posterior distributions, so the three chains of MCMC converge well.

In Bayesian statistics, posterior predictive checkings (PPCs) is a technique used to assess the fit of a statistical model to observed data. It involves using the posterior distributions of the model parameters to generate new datasets and then comparing these simulated datasets with the actual observed data. Here the PPCs results for the BYM CAR model in Figure A.3 are shown as a histogram of column means and standard deviations; on the left is a histogram of column means of lung cancer Y, where the grey bar expresses the distribution of the mean of a particular variable, which is the count of lung cancer, over the simulated dataset, and the vertical red line represents the mean of the observed data for the same variable, and since the red line is located within a large portion of the histogram and falls within the simulated area. Since the red line lies

over most of the histogram and falls near the centre of the distribution of means, the model performs well in capturing concentrated trends in the observed data. The right panel shows the standard deviation histogram, where the vertical red line represents the standard deviation of the count of lung cancer and lies within most of the range of the histogram, indicating that the variability of the model is consistent with that observed in the actual data. Therefore, the results of the PPCs clearly indicate that the model has predictive validity.

CARbym Multiple Years

In the expanded analysis of lung cancer incidence over a decade using the BYM CAR model, we interpret the coefficients within a multivariate context, where each covariate's effect is conditional on the presence of others. This approach allows us to dissect the complex interplay between various risk factors and lung cancer incidence rates. Furthermore, the formula used in the multi-year model resembles that of the single-year model, with the addition of time t as a subscript to the parameters to account for temporal variations.

The `intercept`, valued at -6.6111, serves as a baseline for lung cancer incidence when covariates are at their reference levels. This fixed point is crucial for contextualizing the effects of the other variables within the model. The coefficient for `Smoking Rate` stands at 0.0132, reinforcing the established role of smoking in elevating lung cancer risk. This effect size is interpreted while controlling for the influence of other variables, suggesting that the contribution of smoking to lung cancer remains significant across the timespan even when accounting for additional factors in the model.

With respect to employment, the `Manufacturing` and Construction show negative coefficients of -0.0056 and -0.0119, respectively. This should be interpreted as the effect of manufacturing and construction employment on lung cancer incidence after adjusting for other factors in the model. It is essential to consider this finding in light of the model's structure, where the effect of employment is contingent upon the levels of other covariates being constant. This might suggest the presence of protective mechanisms within the sectors or a proxy for other variables not directly measured in the study.

	Mean	2.5%	97.5%	n.sample	% accept	n.effective	PSRF (upper 95% CI)
(Intercept)	-6.6111	-6.8710	-6.3548	1e+05	46.6	6061.2	1
Smoking_Rate	0.0132	0.0082	0.0183	1e+05	46.6	3728.9	1
Manufacturing	-0.0056	-0.0169	0.0058	1e+05	46.6	5354.6	1
Construction	-0.0119	-0.0296	0.0059	1e+05	46.6	5092.9	1
Emissions	0.0070	0.0038	0.0102	1e+05	46.6	3900.1	1
Earnings	-0.0198	-0.0272	-0.0124	1e+05	46.6	7169.7	1
tau2	0.0245	0.0166	0.0340	1e+05	100.0	7286.7	1
sigma2	0.0025	0.0012	0.0046	1e+05	100.0	2610.0	1

Figure 4.12: Result from running CARbym function (2008 - 2017)

The coefficient for `Emissions` is small but positive at 0.007, indicating a tentative association with lung cancer incidence. This relationship is understood in the context of controlling for other covariates, implying that while emissions contribute to

lung cancer risk, the effect is relatively minor when considered alongside other factors. Nonetheless, this warrants a more detailed examination into the environmental determinants of lung cancer, with an emphasis on disentangling confounding influences.

The `Earnings` coefficient is negative (-0.0198), suggesting that, when other factors are controlled for, regions with higher earnings tend to have lower rates of lung cancer. This shift from the positive coefficient observed in the single-year model to the negative one in the multi-year model resonates with the broader epidemiological literature which generally posits an inverse relationship between socio-economic status and adverse health outcomes. It implies that the positive association observed in the single-year model might be an anomaly or a product of year-specific variations and not reflective of the long-term trend captured over ten years. These aspects are often less visible in single-year snapshots due to short-term fluctuations or data anomalies.

Comparatively, the fact that `tau2` is roughly three times larger than `sigma2` (0.0219 vs. 0.0072) is indicative of the relative importance of spatial correlation in the dataset. This ratio shows that the spatial structure has a more substantial impact on the variance of lung cancer incidence in the model than the unstructured effects. It is a quantitative affirmation that spatially-targeted public health interventions might be more efficacious due to the stronger spatial patterns present in the data.

After performing similar model checks and posterior predictive checks for this multi-year BYM model, the trace plots and density plots as shown in the Figure A.4 showed stable variance and no exhibited trends or drifts, indicating good convergence. Additionally, the Gelman-Rubin test resulted in a value of 1.01, which aligns with the following Figure A.5 and demonstrates that the three chains of this MCMC have converged well. For the PPCs as shown in the Figure A.6, the red line lies over most of the histogram and falls near the centre of the distribution of means, indicating good model performance.

The coefficient for `Smoking Rate` in the single-year model is approximately double the value of the coefficient in the multiple-year model (0.0266 vs. 0.0132). This might suggest that the immediate effect of smoking on lung cancer incidence is more pronounced when looking at data from a single year compared to a broader timespan where long-term factors might dilute the immediate yearly impact. However, in both cases, the positive sign of the coefficient confirms the established understanding that an increased smoking rate is associated with higher lung cancer incidence.

The transformation of the `Earnings` coefficient from positive in the single-year model to negative in the multiple-year model is particularly noteworthy. In the short term, represented by the single-year data, higher earnings have a fairly small positive association with lung cancer rates. This could be influenced by specific economic conditions or health factors relevant to that year, possibly reflecting short-term deviance or confounding variables not captured by the model. In comparison, the negative coefficient for `Earnings` in the multiple-year model aligns with the broader literature. It indicates that over a longer period, regions with higher earnings typically have better health outcomes, including lower lung cancer incidence. The negative value reflects the benefits

of higher socio-economic position, like improved access to healthcare and healthier lifestyle choices, which may contribute to reduced lung cancer rates over time. This shift from positive to negative reasserts the importance of considering temporal effect in understanding the relationships between socio-economic factors and health conditions.

4.3.3 CARleroux Modelling

The alternative CAR prior model, introduced by Leroux et al. in 2000 [18], presents a methodology to capture varying degrees of spatial autocorrelation through a singular set of random effects $\phi = (\phi_1, \ldots, \phi_n)$, which follow a multivariate Gaussian distribution [15]

$$\phi|\mathbf{W}, \tau^2, \rho, \mu \sim \mathcal{N}\left(\mu, \tau^2 \left[\rho\mathbf{W}^* + (1-\rho)\mathbf{I}_n\right]^{-1}\right). \tag{4.5}$$

This prior has a constant non-zero mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, while the precision matrix is given by

$$\mathbf{Q}_L = \rho\mathbf{W}^* + (1-\rho)\mathbf{I}_n,$$

where $\mathbf{I}_n$ represents an $n \times n$ identity matrix and the entries of $\mathbf{W}^*$ are

$$w^*_{jk} = \begin{cases} n_k & \text{if } j = k \\ -1 & \text{if } j \sim k \\ 0 & \text{otherwise.} \end{cases}$$

Leroux model introduces a framework that accommodates spatial correlation ρ in the data [18], which quantifies the strength of the correlation between neighbouring areas in the spatial model. Here, ρ can take on values between 0 and 1. Specifically, the special case $\rho = 0$ implies no spatial autocorrelation, corresponding to independence ($\phi_k \sim \mathcal{N}(0, \tau^2)$). The joint distribution 4.5 is proper if $0 \leq \rho < 1$, while $\rho = 1$ corresponds to the improper intrinsic model (i.e. the BYM CAR model) given by 4.4. The latter indicates that the spatial process is entirely spatially autocorrelated.

$$\begin{aligned} \psi_k &= \phi_k \\ \phi_k|\phi_{-k}, \mathbf{W}, \tau^2, \rho &\sim \mathcal{N}\left(\frac{\rho\sum_{i=1}^K w_{ki}\phi_i}{\rho\sum_{i=1}^K w_{ki} + 1 - \rho}, \frac{\tau^2}{\rho\sum_{i=1}^K w_{ki} + 1 - \rho}\right) \\ \tau^2 &\sim \text{Inverse-Gamma}(a, b) \\ \rho &\sim \text{Uniform}(0, 1). \end{aligned}$$

The formula for the single-year Leroux model shares similar parameter definitions with that of the BYM model. The only difference lies in the expression for ψ_k, which, in the Leroux model, excludes the θ term found in the BYM model. Thus the Leroux model enhances flexibility by using ρ as the criterion of the degree of spatial autocorrelation. When ρ is set to 1, the Leroux model emulates the BYM model, indicating strong spatial autocorrelation. This can be a limitation if the true spatial autocorrelation is less than

perfect. This feature provides the Leroux model with the capacity to modulate the extent of spatial correlation, thereby spanning the spectrum from uncorrelated random effects to a fully spatially correlated framework.

CARleroux Single Year

The Leroux CAR model, fitted to the 2016 dataset, offers insights into the factors influencing lung cancer incidence. The model's `intercept` is estimated at -7.2652, setting a baseline level of lung cancer incidence when the other covariates in the model are at their mean levels. Each covariate's effect is interpreted conditionally, recognising the influence of other variables remaining fixed in the model.

	Mean	2.5%	97.5%	n.sample	% accept	n.effective	PSRF (upper 95% CI)
(Intercept)	-7.2652	-8.1987	-6.3465	1e+05	43.0	4448.3	1
Smoking_Rate	0.0265	0.0067	0.0460	1e+05	43.0	3443.4	1
Manufacturing	0.0070	-0.0275	0.0422	1e+05	43.0	4293.5	1
Construction	-0.0117	-0.0596	0.0342	1e+05	43.0	4935.8	1
Emissions	-0.0005	-0.0123	0.0111	1e+05	43.0	2830.3	1
Earnings	-0.0005	-0.0228	0.0224	1e+05	43.0	6278.7	1
tau2	0.0178	0.0063	0.0392	1e+05	100.0	15871.3	1
rho	0.6143	0.1158	0.9623	1e+05	57.1	20239.7	1

Figure 4.13: Result from running CARleroux function (2016)

The `Smoking Rate`, with a mean estimate of 0.0265, suggests that, holding all other variables constant, there is a positive association between the smoking rate and lung cancer incidence. This is consistent with the fact that smoking is a well-known risk factor for lung cancer. It also underlines the importance of smoking as a persistent and primary risk factor for lung cancer within the studied population.

In contrast, the `Manufacturing` coefficient is slightly negative at -0.0070. This may point to protective aspects within the manufacturing sector or reflect the confounding effects of other variables not captured within the model. The `Construction` employment shows a negative association with a coefficient of -0.0117. This could be due to a range of factors, such as healthier worker populations or effective health and safety regulations in the construction industry. Interestingly, the `Emissions` coefficient is near zero (-0.0005), indicating that, in the context of this model, changes in emissions have a negligible direct effect on lung cancer incidence.

The `Earnings` coefficient, also near zero at -0.0005, presents a contrast to the positive association seen in the single-year BYM CAR model. This reveals that within the specific context of the Leroux model for the year 2016, higher earnings do not significantly increase or decrease lung cancer incidence rates. The shift to a negative coefficient here, although small, aligns with broader literature, proposing that higher socio-economic status often correlates with better health outcomes.

In this model, the mean coefficient for `tau2` is 0.0178, which suggests the variance component of the spatially structured random effects is present, but not overly dominant. This indicates some degree of spatial variation in lung cancer incidence rates across

the council areas, suggesting that geographic factors do play a role, but they might not be the only driving factors. The mean coefficient for `rho` at 0.6143 represents the spatial correlation parameter in the Leroux CAR model. This value is moderately high and suggests that there is a substantial spatial autocorrelation in the lung cancer incidence data. Specifically, this means that council areas that are geographically close to each other are likely to have more similar lung cancer incidence rates than those that are far apart. The implication is that spatially proximate areas may share common environmental or social factors influencing lung cancer rates.

Also considering the model check and posterior predictive checking for this CARLeroux model, the trace plots and the density plots shown in Figure A.7 demonstrated good convergence since there are no obvious multiple peaks. Second, by observing the results of the Gelman-Rubin diagnostic in Figure A.8, the three chains of MCMC converge well. Third, for the results of the posterior predictive checking Figure A.9, the histogram indicates that the observed data statistics, including the mean and standard deviation, fall within the expected range of the posterior predictive distribution. The red line is positioned within most of the histogram's range, but not at the peak, suggesting that the model fits the data reasonably well, but there may be a slight systematic error.

CARleroux Multiple Year

Utilising the Leroux CAR model to analyse lung cancer incidence over a decade (2008 - 2017) allows for an understanding of the role each covariate plays when viewed about the others. This model employs a multivariate framework, where the impact of each factor is considered within the context of the rest. In this model the `intercept` is placed at -6.5180 as with the BYM CAR model, this fixed reference point is essential for interpreting the effects of other variables in the model.

	Mean	2.5%	97.5%	n.sample	% accept	n.effective	PSRF (upper 95% CI)
(Intercept)	-6.5180	-6.7720	-6.2663	1e+05	43.4	5401.2	1
Smoking_Rate	0.0127	0.0079	0.0175	1e+05	43.4	3367.9	1
Manufacturing	-0.0048	-0.0162	0.0063	1e+05	43.4	4240.9	1
Construction	-0.0118	-0.0286	0.0051	1e+05	43.4	4459.8	1
Emissions	0.0068	0.0038	0.0098	1e+05	43.4	2980.2	1
Earnings	-0.0196	-0.0267	-0.0125	1e+05	43.4	8901.2	1
tau2	0.0288	0.0217	0.0374	1e+05	100.0	22098.8	1
rho	0.8522	0.7168	0.9449	1e+05	47.1	28760.2	1

Figure 4.14: Result from running CARleroux function (2008-2017)

The `Smoking Rate` coefficient here is valued at 0.0127. This is somewhat lower than the single-year model, but it still affirms smoking as a significant risk factor for lung cancer incidence. This persistence across the decade emphasises the chronic, enduring impact of smoking on public health, despite any temporal variations.

Regarding employment, both `Manufacturing` and `Construction` exhibit negative coefficients, -0.0048 and -0.0118, respectively. These findings suggest that higher employment in these sectors correlates with lower lung cancer incidence when other factors are held constant. While this could indicate occupational health advancements,

it may also suggest that these variables act as proxies for other unmeasured influences that affect health outcomes. The `Emissions` coefficient, at 0.0068, is positive, indicating a potential association with increasing lung cancer incidence. Although modest, this relationship persists even when adjusting for other covariates.

In stark contrast to the single-year analysis, `Earnings` has a significant negative coefficient of -0.0196. This turnaround from the single-year model supports the consensus that higher socio-economic status, over time, contributes to better health outcomes, including a lower incidence of lung cancer. The negative coefficient in the multi-year analysis reflects the long-term benefits of higher income levels, which may be obscured in a single-year profile due to year-specific anomalies or other confounding factors.

Crucially, the spatial parameters `tau2` and `rho`, at 0.0288 and 0.8522 respectively, reveal substantial spatial dependencies within the lung cancer incidence data. `tau2` denotes the variance of spatially structured random effects, while `rho` measures the proportion of variance explained by the spatial component of the model. The high value of `rho` here is indicative of a strong spatial correlation. This implies that the lung cancer incidence does not occur randomly across space but is instead highly structured and potentially influenced by region-specific factors.

For the model check and PPCs of this large CARLeroux model, the conclusion is similar. The trace and density plots in Figure A.10 indicate good convergence of the sample (which is beta here), and the results of the Gelman-Rubin diagnostic in Figure A.11 demonstrate the convergence of the three chains in the model. Lastly, for the PPCs shown in Figure A.12, the red lines for both plots lie within most of the histogram's range but not at the peak, indicating the possibility of a small systemic error.

4.3.4 Model Comparison

In conducting a comparative analysis of the multi-year models, we here use the Watanabe-Akaike Information Criterion (WAIC) values, with lower values indicating a better balance of model fit and complexity. Table 4.1 demonstrates the evaluation of five different models introduced in the above content.

Overall, it is evidence that the Poisson model registers the highest WAIC value at 3300.071 compared with other multi-year models, suggesting a less favourable fit. This model's performance points to the significance of spatial modelling approaches when analysing data where geographic patterns are presumed to influence the outcomes. Among the multi-year CAR models evaluated, the WAIC values for BYM and Leroux are notably similar, with a marginal difference of 0.2, suggesting a comparable fit between these spatial approaches. Then when considering the single-year models, the BYM CAR model exhibits a slightly better WAIC value of 254.893, compared to Leroux's 256.609. This points towards a potential advantage for the BYM model in terms of out-of-sample predictive accuracy.

These findings emphasise the critical role that spatial relationships play in the analysis of lung cancer incidence data. The improved fit of the spatial models, especially the BYM

Model	WAIC
BYM (Single Year)	254.893
Leroux (Single Year)	256.609
BYM (All Year)	2572.636
Leroux (All Year)	2579.836
Poisson (All Year)	3300.071

Table 4.1: WAIC Values for Various Models

model, underscores the potential influence of geographic factors on disease incidence. It also stresses the necessity of integrating spatial dependencies to adequately capture the complexities of the data, which non-spatial models, such as the Poisson model, might fail to detect. Therefore, for studies where geographic correlation is hypothesized to be influential, spatial models, particularly the CAR BYM model, should be given precedence, which will be discussed in Chapter 5.

Chapter 5

Prediction

5.1 Regression Prediction

Our research meticulously applies Poisson Regression analyses as discussed in the Chapter 3, leveraging data spanning from 2008 to 2016, with the objective of projecting lung cancer rates for the year 2017. This analytical approach allows us to assess the influence of various covariates on lung cancer incidence, offering nuanced insights into the dynamics at play. Equation 5.1 delineates the predictive formula, where $\hat{Y}_i$ denotes the predicted number of lung cancer cases across 32 council areas, and Pop_count_i corresponds to the population counts. The remaining parameters retain the meanings attributed earlier in the report.

$$\hat{Y}_i = \exp\left(\underline{\beta}^{\mathsf{T}} \underline{X}_i\right) \cdot \text{Pop\_count}_i \tag{5.1}$$

5.2 Bayesian Model Prediction

This section focuses on the prediction of the lung cancer cases based on the two CAR models (BYM & Leroux) implemented in Chapter 4. Consistent with before, the prediction will include both a single-year approach to 2016 and a multi-year approach from 2008 to 2016. In the single-year 2016 analysis, the model's predictive capacity is based on the covariate relationships and lung cancer incidence observed specifically for that year. While, the multi-year model analysis extends the CAR model's application to encompass a broader temporal scope, aggregating data from 2008 to 2016.

5.2.1 CARbym

The equation 5.2 presents the predictive formula for estimating the lung cancer incidence in the year 2017, utilizing the beta coefficients from the model outlined in Chapter 4. In this equation, $\hat{\mathbf{Y}}$ represents the anticipated number of lung cancer cases in the 32 council areas, while α is the intercept term. The vector β represents the posterior mean across five different covariates and $\mathbf{X}$ is the covariate matrix. It is noteworthy that while both the single-year and multi-year models employed the same formula but substitute with different size of the matrix. For the single-year model, the dimension of all the

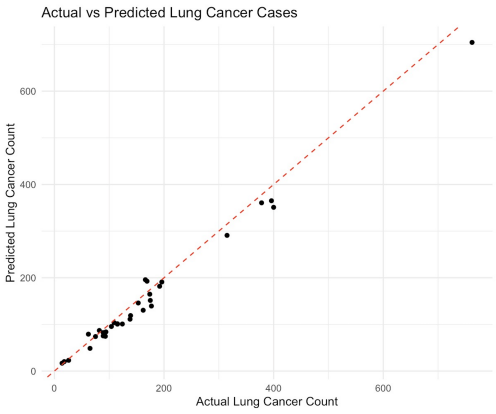


Figure 5.1: Single year

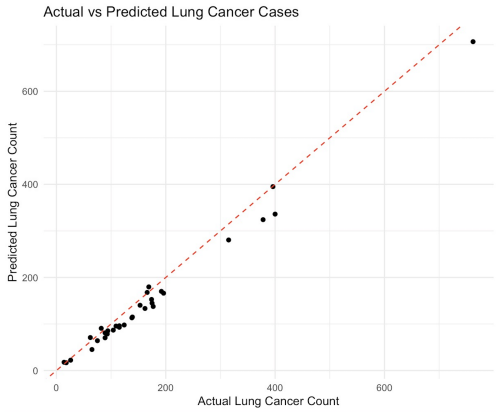


Figure 5.2: Multiple year

elements are the same with Equation 4.3 when $t = 1$; For the multi-year model, the dimensions are all the same with Equation 4.3 when $t = 9$.

$$\hat{\mathbf{Y}} = \boldsymbol{\alpha} + \mathbf{X}^{\mathbf{T}}\boldsymbol{\beta} \tag{5.2}$$

The Figures 5.1 and 5.2 can visualize the difference between Actual and Predicted Lung cancer counts, since the red line represents the line of perfect prediction, where the predicted values are exactly equal to the actual values. Both the single-year and multiple-year prediction plots exhibit a similar trend in the distribution of points: the models tend to accurately predict the actual lung cancer cases for lower incidence rates, as evidenced by the clustering of points around the line of perfect prediction at the lower end of the count scale. However, as the actual number of cases increases, both models show a tendency to underpredict — the points deviate above the line of perfect prediction, indicating the predicted values fall short of the actual counts.

5.2.2 CARleroux

The equation 5.2 also outlined the predictive formula for Leroux CAR model, and share the same size of the matrix for single-year and multi-year with BYM models. The distinction is the random effect $\phi_{\text{mean},i}$ is the mean of the spatial random effects for the i-th council area, averaged over all MCMC chains. Unlike the BYM model, the Leroux model does not provide estimates for ψ, because the MCMC samples do not yield an estimate for this parameter. Thus it might not be a good choice to compare the prediction effect of BYM and Leroux model, as the Leroux is built based on the BYM, we still tend to do this as to apply. While it may not be strictly appropriate to compare the predictive efficacy of the BYM and Leroux models, due to the fact that the latter's derivation is from the former, such a comparison is nonetheless undertaken to explore the practical application and effectiveness of both modelling approaches in our analysis.

Similar to the observations made earlier, the prediction plots 5.3 and 5.4 for the two Leroux models exhibit remarkable similarity, signifying that both models are equivalently effective in capturing the underlying spatial patterns and associations with lung

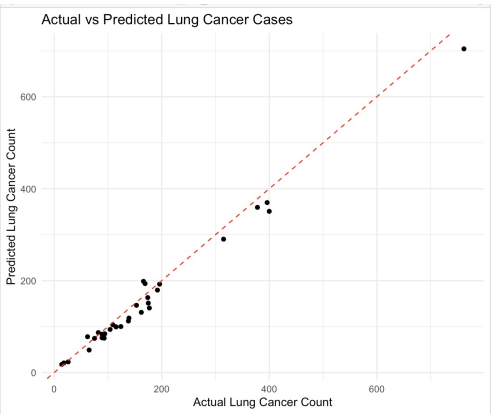


Figure 5.3: Single year

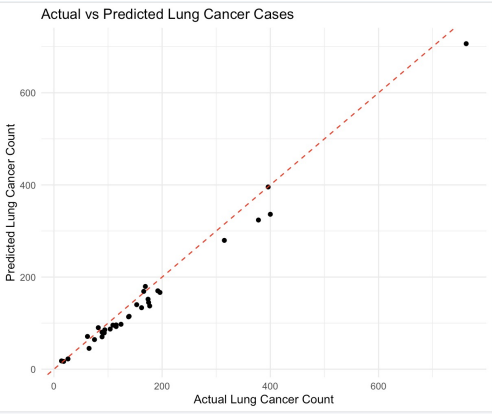


Figure 5.4: Multiple year

cancer incidence. Furthermore, the overall trend in these plots closely mirrors that of the BYM model plots previously discussed, indicating identical model performance.

5.2.3 Model Comparison

Root Mean Squared Error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed from the environment that is being modelled or estimated. RMSE is a standard way to measure the error of a model in predicting quantitative data. Formally, the RMSE is the square root of the mean squared error (MSE), which is the average squared difference between the estimated values and the actual value. The RMSE is defined by the formula:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}.$$

where n is the number of observations, Y_i is the observed values for the ith council area, and $\hat{Y}_i$ represents the predicted values. It is considered in the form of a percentage.

Furthermore, the Root Mean Square Error (RMSE) is widely recognized for assessing the predictive accuracy of models. However, in our analysis, an outlier evident in the graphs above away significantly from the cluster of data points, potentially impacting the overall RMSE calculation. To account for this and provide a more balanced evaluation of the model's performance, we introduce an additional criterion, the Root Mean Square Percentage Error (RMSPE).

The root mean square percentage error (RMSPE) of the model is defined as a measure of prediction accuracy in a regression model, expressed as a percentage. It provides an idea of the relative error between the predicted and observed values, making it particularly useful when we want to understand the error in terms of a percentage of the actual values due to some large-scale point that may affect our overall MSE.

The formula to calculate RMSPE is:

$$\text{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{Y_i}\right)^2} \times 100\%,$$

Model	RMSE	RMSPE
Bym (single-year)	22.0945	13.9245
Leroux (single-year)	22.2243	14.3121
Bym (multi-years)	25.6103	15.6510
Leroux (multi-years)	25.4872	15.5554
Poisson	36.2129	19.5983

Table 5.1: RMSE, and RMSPE Values for Various Models

The values displayed in Table 5.1 show the comparison of predictive efficacy across different spatial models. By the nature of RMSE and RMSPE, lower values are indicative of a more accurate model. It is evident that the Poisson model yields the highest RMSE and RMSPE, suggesting lower prediction accuracy relative to the multi-year CAR models, as the non-spatial model does not have a random effect or parameter to consider. Moreover, the Leroux model with multi-years owns the smallest values among multi-years models, indicating the superiority compared with BYM multi-years model. This advantage could be attributed to the Leroux model's dynamic adaptation of the random effect for each time period, as opposed to the BYM model which assumes a fixed random effect ϕ. Then comparing the CAR model for the single-year, the BYM model has a smaller value than the Leroux model, the reason may be that the variability of the random effect does not show its benefit as it does in the multi-years model.

Furthermore, multi-years models yield higher values than single-year models. However, the RMSE of the multi-years models is only slightly higher by about 3 units compared to the single-year models, indicating that extending the time frame of data does not significantly compromise prediction accuracy. Based on the available data, the best prediction can be achieved by modelling with 2016 data to predict 2017 data, likely because variables like emissions and earnings do not fluctuate drastically year-over-year. However, using the previous year's data to predict the following year may lead to more accurate results. As mentioned in the previous chapter of the DATA, lung cancer rates have not shown a clear trend over time for most regions. Therefore, if the dataset covers a longer period, a multi-year model can provide accurate predictions.

Chapter 6

Reality Problems Solving

Spatial autocorrelation analysis and Bayesian spatial modelling, including CAR models, have established smoking as the most critical factor influencing lung cancer incidence, with income levels also identified as a significant determinant. The analysis demonstrates a strong positive relationship between smoking rates and lung cancer diagnoses, with high-smoking areas typically exhibiting elevated lung cancer rates. Conversely, regions with higher incomes display significantly lower rates of lung cancer, suggesting a negative association between income and lung cancer incidence.

The outcomes emphasises the complex interplay of socioeconomic factors and health behaviors. The strong correlation between smoking prevalence and lung cancer rate accentuates the urgent needs for effective smoking cessation programmes and targeted public health campaigns. These initiatives are crucial in mitigating tobacco use across various population strata. At the same time, the noticeable impact of socio-economic status on health outcomes requires the formulation of policies that address income disparities to reduce lung cancer risks effectively.

This chapter will investigate these findings, with the discussion of implications of the relationships among lung cancer, smoking, and earning. Accordingly, potential policy solutions will also be explored to address these challenges. From the comparison table 4.1, we have already found that BYM (All Year) is the model with the lowest WAIC, hence we use the estimate of it from Figure 4.12 as a foundation for the further investigation.

6.1 Smoking and Socio-economic Factors

The examination of Local Indicators of Spatial Association (LISA) Cluster Plots 4.10 has illustrated three council areas - Inverclyde, Renfrewshire, and West Dunbartonshire - as demonstrating High-High clusters. This denotes regions where areas exhibiting higher incidences of lung cancer are encircled by areas with similarly high incidences. Such a configuration, indicative of positive spatial autocorrelation, clarifies the existence of “hotspots” wherein lung cancer rates are consistently raised across adjacent areas. These council areas, contiguous with the western neighbour of Glasgow City, highlight a pronounced aggregation of lung cancer prevalence within this area.

Exploring deeper into socio-economic indicators, income distribution data disseminated by the Scottish Government [29] for the year 2018 further enriches this analysis. The median per capita weekly incomes in Inverclyde, Renfrewshire, and West Dunbartonshire are recorded at £450, £540, and £480, respectively, each notably below the Scottish average of £550. This financial delineation accentuates the inverse correlation between income levels and lung cancer rates within these areas, providing a nuanced understanding of the socio-economic backdrop against which these health patterns emerge. The confluence of spatial epidemiological findings with detailed income data not only corroborates the initial observations but also offers a more granular perspective on the interplay between economic status and lung cancer incidences.

Here, two council areas identified as High-High clusters (Renfrewshire and West Dunbartonshire) and two as Low-High clusters (East Renfrewshire and East Dunbartonshire) all reside within the Greater Glasgow boundaries, reflecting a socio-economic divide in smoking prevalence and its consequent health implications.

In contrast to the western council areas, East Renfrewshire and East Dunbartonshire, which locate in the eastern part of Great Glasgow, are characterized as Low-High areas, as depicted in light blue on the cluster maps. These areas demonstrate low lung cancer rates despite being surrounded by locations with higher incidences of the disease. This unique positioning may classify them as spatial outliers, indicative of negative spatial autocorrelation, and suggests that there may be distinct underlying factors contributing to the lower lung cancer rates in these areas, divergent from those influencing the neighboring regions. Consistent with predictions, East Renfrewshire and East Dunbartonshire demonstrate considerably higher income levels. The median per capita weekly incomes for these areas are £740 and £720, respectively, significantly surpassing those observed in the aforementioned western council areas adjacent to Glasgow. This economic prosperity further emphasises the relationship between higher income levels and lower incidences of lung cancer.

From Figure 6.1, the location of the dots within the violin plot distributions indicates key similarities between the two council areas. Notably, the dots representing earnings are situated towards the upper portion of the distribution for both areas, suggesting that both have relatively high-income levels when compared to the smoking rate and lung cancer ratio. This upper placement of the earnings dots could imply that the populations in these council areas are generally well-off financially. The plot for the smoking rate shows a relatively narrow and bottom-heavy distribution for both council areas, which indicates that smoking rates tend to be on the lower end across the years. The bulk of the points clustering toward the narrower part of the plot may suggest a consistent control or reduction in smoking prevalence over the observed period. This could reflect effective anti-smoking policies or health awareness programs in these areas. The distribution's tapered shape at the top, however, with fewer points, signals that there were years or areas where the smoking rate was higher, although these appear to be exceptions rather than the rule.

The clustering of data points towards the middle and lower sections of the lung cancer

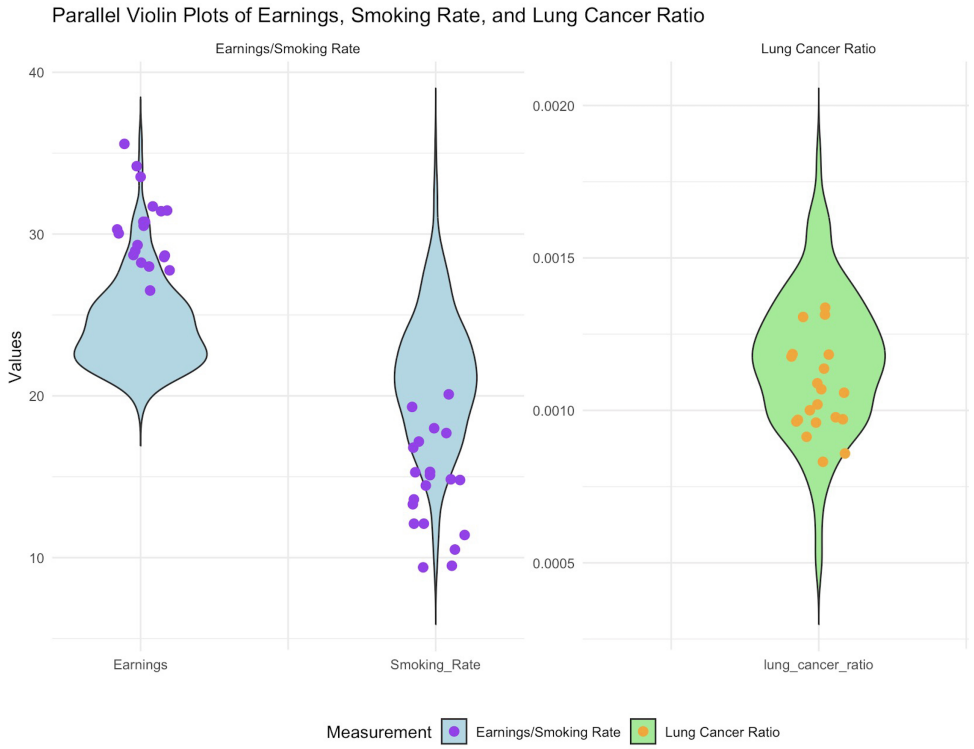


Figure 6.1: Violin plot for Earnings, Smoking Rate and Lung Cancer Ratio in East Renfrewshire and East Dunbartonshire

ratio's violin plot indicates that for most of the ten years, both council areas experienced moderate to low lung cancer rates. This aligns with public health expectations, as these areas also exhibit higher earnings and lower smoking rates. The combination of higher socioeconomic status and lower smoking prevalence is typically correlated with better health outcomes, which seems to be reflected in the lower incidence of lung cancer. The presence of some points towards the top-middle, however, suggests occasional increases in lung cancer rates, which could be attributable to factors other than smoking or earnings, such as genetic predispositions, localized environmental factors, or lifestyle choices not captured by the smoking rate. It could also reflect the natural variation in cancer incidence rates year over year.

The overall picture presented by the violin plots indicates a scenario that is consistent with established health determinants: areas with higher earnings and lower smoking rates tend to have lower lung cancer ratios. Yet, the outliers or points that deviate from this general trend hint at the complexity of cancer epidemiology, where multiple factors—including those not directly measured in this analysis—can influence health outcomes. This visual representation underscores the importance of continued public health surveillance and multifactorial analysis to identify and address all possible contributors to lung cancer within a population.

The analysis reveals a negative correlation between smoking and income, with data suggesting that smokers tend to have lower earnings compared to non-smokers, a phenomenon supported by empirical evidence from Reed's article. Smokers' weekly

earnings are on average 6.8% lower than those of non-smokers, which is considered an earnings penalty for smoking (Reed, 2020[25]). This trend is paralleled by the observation that lower socio-economic groups often have higher smoking rates, a situation exemplified in the Greater Glasgow area. Gray and Leyland [9] claims that the Glasgow area has higher smoking rates than the rest of Scotland due to both individual and local socioeconomic situations, particularly among men. This reflects its poorer socio-economic position. In fact, both council areas have a smoking prevalence rate of 8.6% in 2019, which is the lowest among these 32 council areas, indicating consistency with our findings.

6.2 Policy

6.2.1 Smoking Cessation

In light of the strong association between smoking and lung cancer incidence, as well as the negative correlation between income levels and smoking rates, it is imperative to develop a comprehensive suite of policy solutions aimed at smoking cessation. These policies must be multifaceted and synergistic, addressing the issue through public health education, regulatory measures, and healthcare interventions. To effectively reduce the burden of lung cancer, particularly in lower-income populations where smoking prevalence is higher, policy initiatives must prioritise accessibility to cessation programs, enforce anti-smoking legislation, and foster environments conducive to quitting. Addressing the nexus between smoking and lung cancer through policy intervention primarily hinges on effective smoking cessation strategies. While fostering economic growth to raise earnings is an enduring challenge within the field of economics, a more direct and immediate impact on lung cancer incidence can be made through targeted tobacco control measures. Additionally, these measures should be supported by robust research and surveillance to monitor their effectiveness and inform continuous improvement.

The literature currently in publication highlights the efficacy of a dual strategy combining pharmaceutical and behavioural therapies in the fight against lung cancer through smoking cessation. It has been shown that behavioural and psychological therapies, including as cognitive behavioural therapy (CBT), motivational interviewing, acceptance and commitment therapy, behavioural therapy, and contingency management, considerably increase the success rates of cessation. By giving people tools to control cravings, withdrawal symptoms, and triggers, these approaches tackle the multifaceted nature of nicotine addiction and promote a more sustainable cessation process (United States Public Health Service Office of the Surgeon General; National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health, 2020[32]).

Furthermore, substantial research has been implemented on Nicotine Replacement Therapy (NRT), which comes in a variety of forms such as gums, patches, inhalers, nasal sprays, and lozenges. Research indicates that NRT is relatively effective in aiding smoking cessation when compared to placebo or control interventions, with efficacy evaluated at different time frames post-cessation efforts (Wu, P., Wilson, K., Dimoulas,

Type of Treatment (Reference)	Quit Rate			Number Needed to Treat
	Placebo	Treatment	Difference	
	Percentage			
Nicotine gum, 2 mg	20	38	18	6
Nicotine gum, 4 mg	20	43	23	4
Nicotine inhaler	8	13	5	20
Nicotine nasal spray	10	26	16	6
Nicotine patch	12	19	7	14
Sustained-release bupropion hydrochloride	12	23	11	9

Figure 6.2: Approximate number needed to treat per patient who quits successfully[12]

P. *et al.*, 2006[34]). This suggests that NRT can play an important role in the initial stages of quitting by relieving withdrawal symptoms and reducing the urge to smoke.

As shown in Karnath's study (2002)[12], the effectiveness of various smoking cessation treatments can be quantified in terms of quit rates and the number needed to treat, which provides a clear comparison of the efficacy of different pharmacotherapies. From Figure 6.2, it can be illustrated that the quit rate for the 4 mg nicotine gum is the highest at 43%, making it the most effective treatment listed. This treatment has a Number Needed to Treat (NNT) of 4, which is the average number needed to treat for patient who quits successfully, indicating high efficiency. The nicotine patch has a moderate quit rate of 19% with an NNT of 14, while sustained-release bupropion hydrochloride presents a quit rate of 23% and an NNT of 9. The nicotine inhaler, although with the lowest increase in quit rate (5%) from placebo, has an NNT of 20, suggesting it may be less effective compared to the other treatments.

6.2.2 Development of Tobacco Regulation

In Scotland, a range of measures and resources have been implemented to support tobacco regulations and smoking cessation efforts. The purpose of such efforts is to offer recommendations to advisers and medical professionals who assist smokers in quitting. These include the provision of specialist services, brief interventions, and harm reduction strategies.

The development of Tobacco Control Policy in Scotland has been marked by significant efforts and strategic plans aimed at reducing tobacco use and achieving a tobacco-free generation by 2034. A qualitative study highlighted the main elements that contributed to the successes of Scotland's tobacco control policies, including strong political leadership, comprehensive mass media campaigns, and robust legislation aimed at reducing

the availability and marketing of tobacco products, as well as exposure to second-hand smoke (Laird *et al.*, 2019[14]). Future policy actions suggested by experts involve focusing on the price and availability of tobacco products, maintaining strong political leadership, building on successful mass media campaigns, and developing methods to address inequalities in smoking prevalence.

To sum up, the strong evidence linking smoking to lung cancer and the negative correlation between income and smoking rates highlight the vital need for an all-encompassing, diversified approach to smoking cessation strategy. Effective policies must intertwine public health education, regulatory frameworks, and healthcare interventions to address this pressing public health issue. Particularly in lower-income communities where smoking rate is higher, policymakers need to ensure that cessation programmes are accessible. Meanwhile, anti-smoking laws are required to be strictly enforced, and supportive environments for quitting should be cultivated. The heart of these policy interventions lies in successful smoking cessation strategies, which can significantly impact lung cancer rates more directly and swiftly than long-term economic growth strategies.

Chapter 7

Conclusion

In the first place, the comparison of WAIC values between multi-years CAR model and Poisson regression model, highlighted the fact that the fluctuations in lung cancer rates influenced by geographical factors. Then, checking the covariate coefficients from both single-year and multi-year CAR models, the influence of the `Manufacturing` and `Emissions` are minimal. In contrast, the coefficient value for `Construction` is associated with decreased lung cancer rates, suggesting potential protective effects. Notably, the coefficient for `Smoking Rate` remains large across different models, affirming their roles as a significant risk factor. The coefficient of `Earnings` is special, as the data in the multi-year model is always larger than the data in the single-year model. This appearance may demonstrate the problem of the single-year model, as it may include the unique circumstances of the year.

After conducting the model comparison for both single-year and multi-year models, the single-year BYM CAR model seems to be the most ideal one as it has the superior value of WAIC, RMSE and RMSPE. Nonetheless, it is not prudent to conclude that the multi-year models possess poor performance: the RMSE values does not show significant difference between the two models. As a matter of fact, the multi-year model contains a broader temporal scope with more corresponding information. Therefore the model should be tested with the data spanning extended time periods.

By performing precise spatial analysis and disease mapping within Scotland, our research not only depicts the geographical distribution of lung cancer incidence but also uncovers the interrelation between smoking rates, socio-economic factors, and lung cancer prevalence. Through spatial autocorrelation analysis and Bayesian spatial modelling, including CAR models, we have demonstrated the profound impact of smoking as the primary determinant of lung cancer rates, with socio-economic position also emerging as a crucial factor. Our findings imply the critical need for targeted public health interventions, particularly smoking cessation programmes. They are not only comprehensive but also tailored to address the specific needs of various population strata, especially in lower-income communities where smoking prevalence is higher.

Moreover, by differentiating our analysis between ”Single Year” and ”Multiple Year” models, we have illuminated the temporal trend of lung cancer incidence, which reveals both annual variations and longer-term trends. While our research is rooted in the

Scottish context, the underlying principles and analytical techniques are universally applicable, providing a blueprint for public health professionals worldwide to investigate and mitigate the burden of lung cancer. These insights are fundamental for developing effective, evidence-based public health strategies and policies aimed at lowering lung cancer incidence through the reduction of smoking rates and addressing socio-economic disparities. In sum, our report provides a comprehensive framework for informing and guiding public health interventions, with the potential to positively impact lung cancer prevalence and improve overall public health outcomes.

Bibliography

[1] J. Besag, J. York, and A. Mollie. "Bayesian Image Restoration with Two Applications in Spatial Statistics". In: *Annals of the Institute of Statistical Mathematics* 43 (1991), pp. 1–59.

[2] Stef van Buuren. *Multivariate Imputation by Chained Equations*. https://cran.r-project.org/web/packages/mice/mice.pdf. Accessed: 23 October 2023. 2023-05-24.

[3] A.J. Cohen. *Air Pollution and Lung Cancer*. National Institutes of Health. Accessed: 15 October 2023. no date. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1746537/pdf/v058p01010.pdf.

[4] Earl Duncan et al. *Developing a Cancer Atlas using Bayesian Methods: A Practical Guide for Application and Interpretation*. Accessed: 05 March 2024. Aug. 2020. URL: https://atlas.cancer.org.au/developing-a-cancer-atlas/index.html (visited on 10/23/2023).

[5] Craig K. Enders. *Applied Missing Data Analysis*. New York: The Guilford Press, 2010. ISBN: 9781606236413.

[6] Environmental Systems Research Institute (Esri). *What is a shapefile?* https://desktop.arcgis.com/en/arcmap/latest/manage-data/shapefiles/what-is-a-shapefile.htm. Accessed: 26 February 2024. 2021.

[7] Dani Gamerman and Hedibert F. Lopes. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. 2nd ed. CRC Press LLC, 2006. URL: https://ebookcentral.proquest.com/lib/ed/detail.action?docID=5379188.

[8] R. C. 1954. Geary. *The Contiguity Ratio and Statistical Mapping*. https://doi.org/https://doi.org/10.2307/2986645. Accessed: 12 Dec 2023.

[9] Linsay Gray and Alastair H Leyland. "Is the "Glasgow effect" of cigarette smoking explained by socio-economic status?: a multilevel analysis". In: *BMC Public Health* 9 (2009), p. 245. DOI: 10.1186/1471-2458-9-245. URL: https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-9-245.

[10] A. Ralph Henderson. *The bootstrap: A technique for data-driven statistics. Using computer-intensive analyses to explore experimental data*. https://atlas.cancer.org.au/developing-a-cancer-atlas/index.html. Accessed: 23 October 2023. 3 June 2005.

[11] Andrew T Jebb et al. "Time series analysis for psychological research: examining and forecasting change". In: *Frontiers in Psychology* 6 (2015), p. 727. DOI: 10.3389/fpsyg.2015.00727. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00727/full.

[12] B. Karnath. "Smoking Cessation". In: *The American Journal of Medicine* (2002). Accessed: 04 March 2024. URL: https://www.sciencedirect.com/science/article/pii/S0002934301011263#aep-abstract-id4.

[13] Tom Koch. *Cartographies of Disease: Maps, Mapping, and Medicine*. Redlands, CA: ESRI Press, 2005.

[14] Yvonne Laird et al. "Tobacco Control Policy in Scotland: A Qualitative Study of Expert Views on Successes, Challenges and Future Actions". In: *International Journal of Environmental Research and Public Health* 16.15 (2019), p. 2659. DOI: 10.3390/ijerph16152659. URL: https://www.mdpi.com/1660-4601/16/15/2659.

[15] Duncan Lee. "A comparison of conditional autoregressive models used in Bayesian disease mapping". In: *Spatial and Spatio-temporal Epidemiology* 2.1 (2011), pp. 79–89. DOI: 10.1016/j.sste.2011.03.001. URL: https://www.sciencedirect.com/science/article/pii/S1877584511000049.

[16] Duncan Lee. *CARBayes version 6.1: An R Package for Spatial Areal Unit Modelling with Conditional Autoregressive Priors*. Tech. rep. Updated version of a paper in the Journal of Statistical Software, 2013, Volume 55, Issue 13. University of Glasgow, 2020.

[17] Duncan Lee. "CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors". In: *Journal of Statistical Software* 55.13 (2013), pp. 1–24. DOI: 10.18637/jss.v055.i13. URL: https://www.jstatsoft.org/htaccess.php?volume=55&type=i&issue=13.

[18] B. Leroux, X. Lei, and N. Breslow. *Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence*. Ed. by M. Halloran and D. Berry. New York, 2000.

[19] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2002. ISBN: 1119013569.

[20] Iris Eekhout Martign W Heymans. *Applied Missing Data Anlaysis: Rubin's Rule*. https://bookdown.org/mwheymans/bookmi/rubins-rules.html. Accessed: 05 March 2024. 2019-01-20.

[21] Paula Moraga. *Spatial Statistics for Data Science: Theory and Practice with R*. New York: Chapman Hall/CRC Data Science Series, 2023.

[22] P. A. P. 1950. Moran. *Notes on Continuous Stochastic Phenomena*. https://doi.org/https://doi.org/10.2307/2332142. Accessed: 12 Dec 2023.

[23] J.S. Neuberger and R.W. Field. *Occupation and Lung Cancer in Non-smokers*. https://www.degruyter.com/document/doi/10.1515/REVEH.2003.18.4.251/html. Accessed: 16 October 2023. 2003.

[24] Edzer Pebesma and Roger Bivand. *Spatial Data Science: With Applications in R*. Boca Raton, FL: CRC Press, 2023. ISBN: 978-0-429-45901-6. DOI: 10.1201/9780429459016.

[25] Howard Reed. *The impact of smoking history on employment prospects, earnings, and productivity: an analysis using UK panel data*. Tech. rep. Report for ASH. Landman Economics, Sept. 2020. URL: https://ash.org.uk/information-and-resources/reports-submissions/reports/smoking-employment-earnings/.

[26] Paul Roback and Julie Legler. *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R*. 1st ed. Accessed: 05 March 2024. Bookdown, 2021. URL: https://bookdown.org/roback/bookdown-BeyondMLR/ (visited on 10/23/2023).

[27] Public Health Scotland. *Cancer incidence in Scotland to December 2021*. Accessed: [05 March 2024]. 2023. URL: https://publichealthscotland.scot/publications/cancer-incidence-in-scotland/cancer-incidence-in-scotland-to-december-2021/.

[28] Public Health Scotland. *Cancer mortality in Scotland - Annual update to 2020*. Accessed: [05 March 2024]. 2021. URL: https://publichealthscotland.scot/publications/cancer-mortality/cancer-mortality-in-scotland-annual-update-to-2020/.

[29] Scottish Government. *Income Distribution Statistics 2018*. https://www.gov.scot/publications/income-distribution-statistics-2018/. Accessed: 05 March 2024. 2018.

[30] MBA Skool Team. *Offset - Definition Meaning*. https://www.mbaskool.com/business-concepts/statistics/7554-offset.html. Accessed: 12 Dec 2023.

[31] W. R. 1970. Tobler. *A Computer Movie Simulating Urban Growth in the Detroit Region*. https://doi.org/https://doi.org/10.2307/143141. Accessed: 23 Nov 2023.

[32] United States Public Health Service Office of the Surgeon General and National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. *Smoking Cessation: A Report of the Surgeon General*. Interventions for Smoking Cessation and Treatments for Nicotine Dependence. US Department of Health and Human Services. 2020. URL: https://www.ncbi.nlm.nih.gov/books/NBK555596/.

[33] J. Wakefield. "Disease Mapping and Spatial Regression with Count Data". In: *Biostatistics* 8.2 (2006). Accessed: 04 February 2024, pp. 158–183. URL: https://academic.oup.com/biostatistics/article/8/2/158/230741 (visited on 02/04/2024).

[34] Ping Wu et al. "Effectiveness of smoking cessation therapies: a systematic review and meta-analysis". In: *BMC Public Health* 6.300 (2006). DOI: 10.1186/1471-2458-6-300. URL: https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-6-300.

[35] Guangbiao Zhou. "Tobacco, air pollution, environmental carcinogenesis, and thoughts on conquering strategies of lung cancer". In: (Nov, 2019). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6936241/.

Appendix A

Corresponding Figures

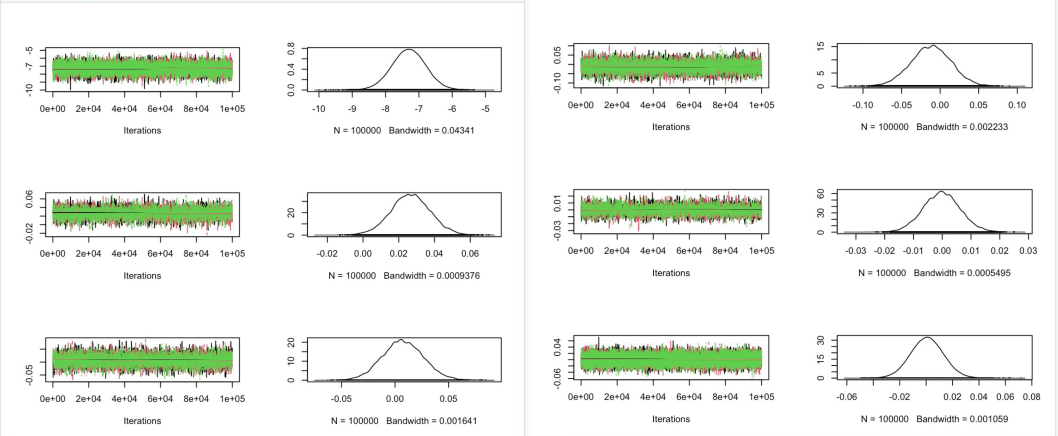


Figure A.1: Converge for CAR BYM

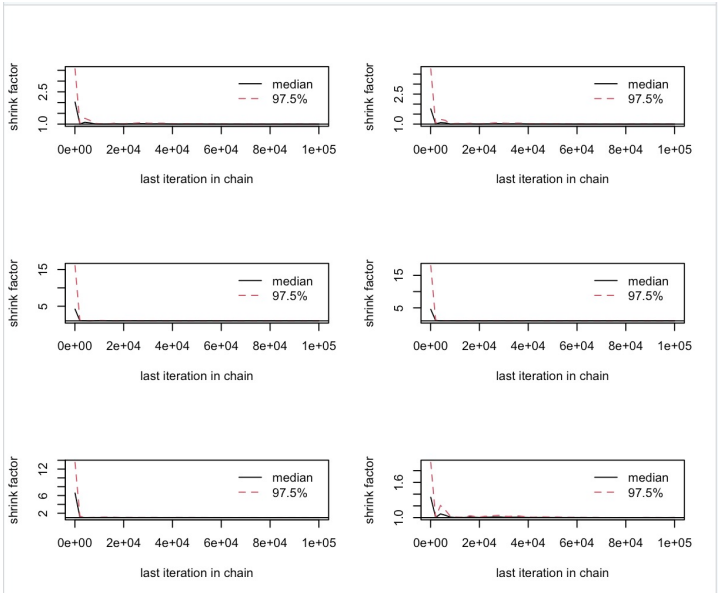


Figure A.2: Gelman-Rubin for CAR BYM

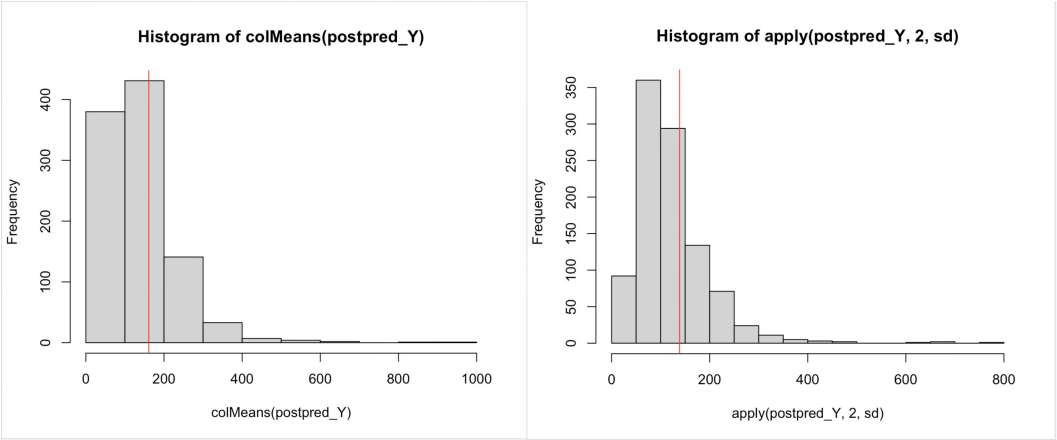


Figure A.3: histogram for CAR BYM

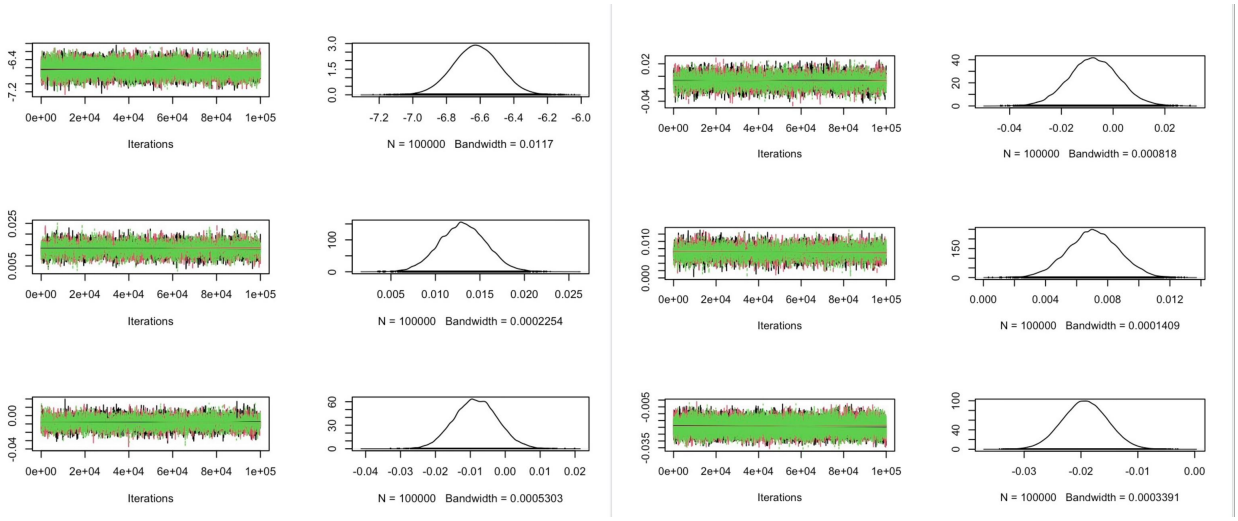


Figure A.4: Converge for CAR BYM big

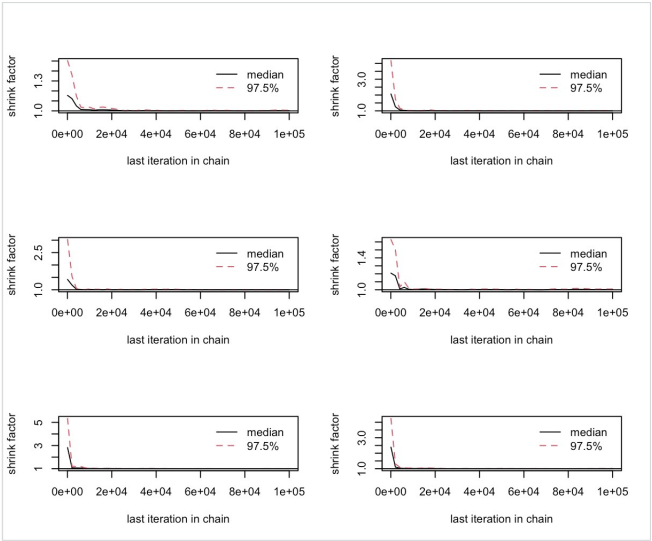


Figure A.5: Gelman-Rubin for CAR BYM big

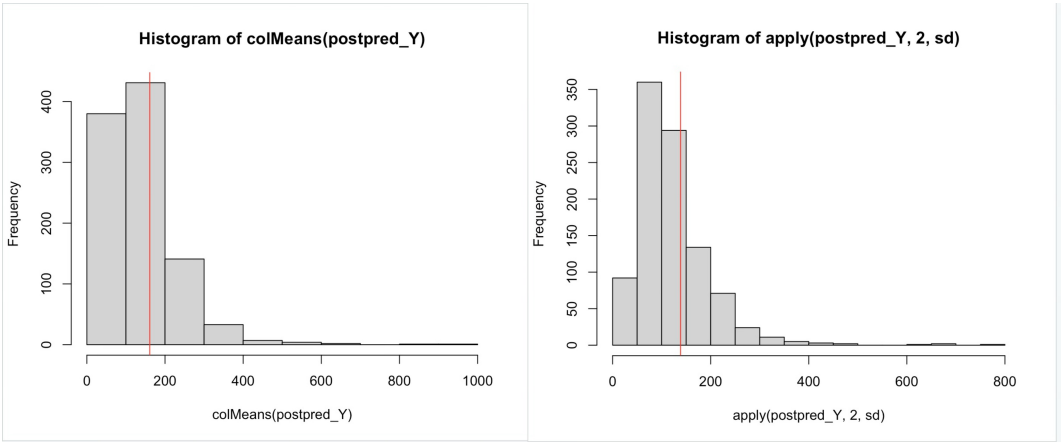


Figure A.6: histogram for CAR BYM big

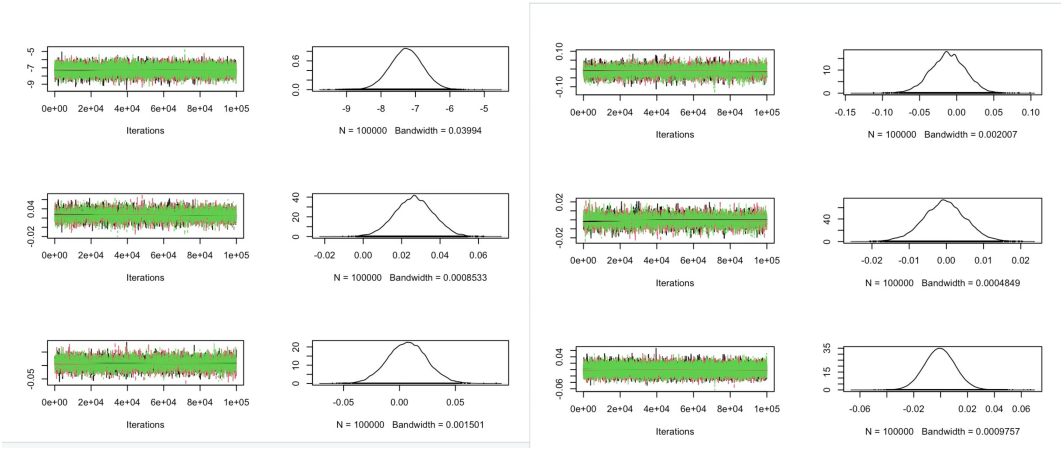


Figure A.7: Converge for CAR LEROUX

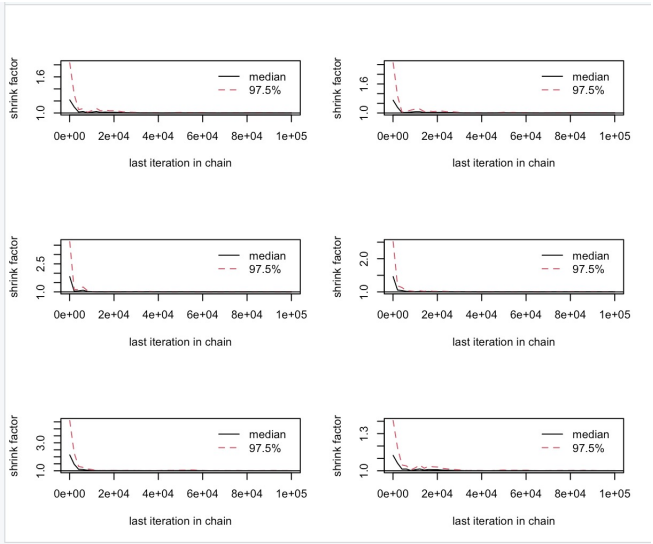


Figure A.8: Gelman-Rubin for CAR LEROUX

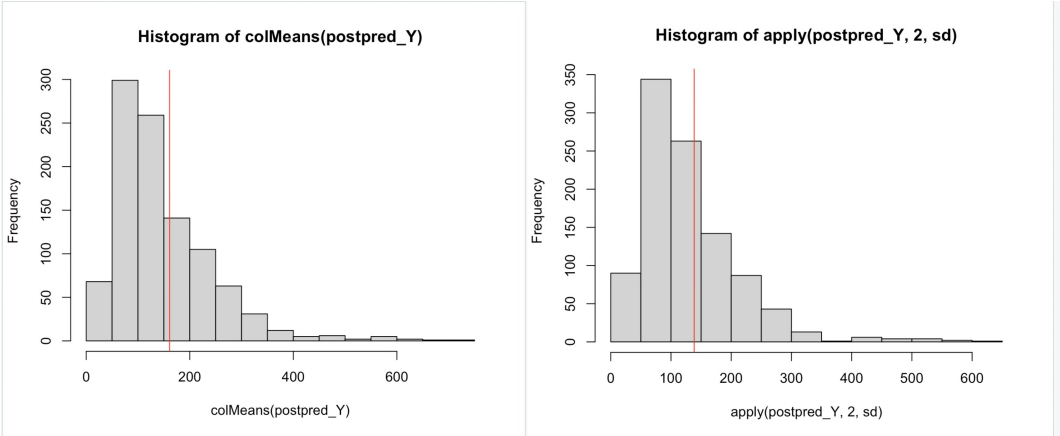


Figure A.9: histogram for CAR LEROUX

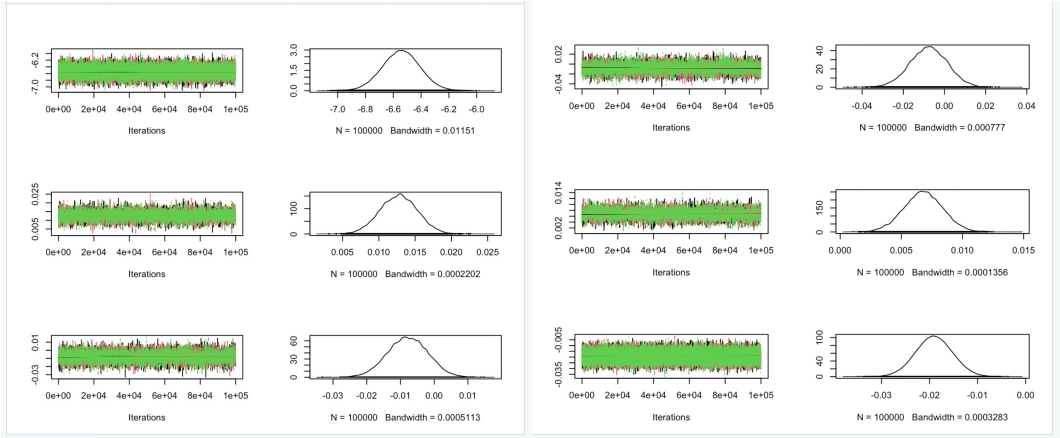


Figure A.10: Converge for CAR LEROUX big

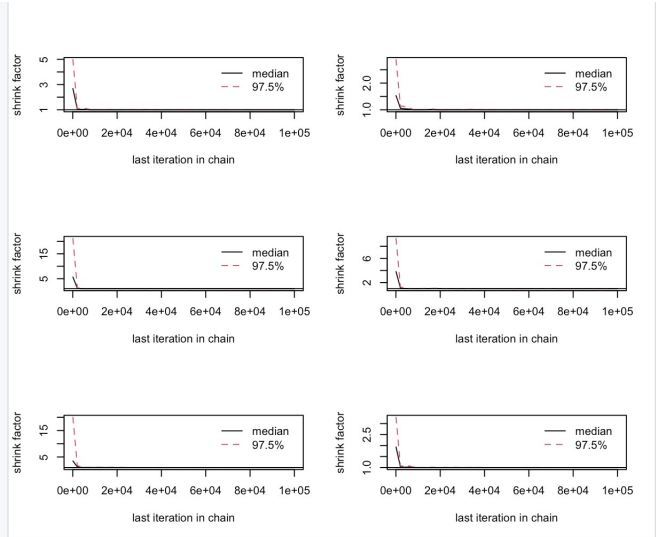


Figure A.11: Gelman-Rubin for CAR LEROUX big

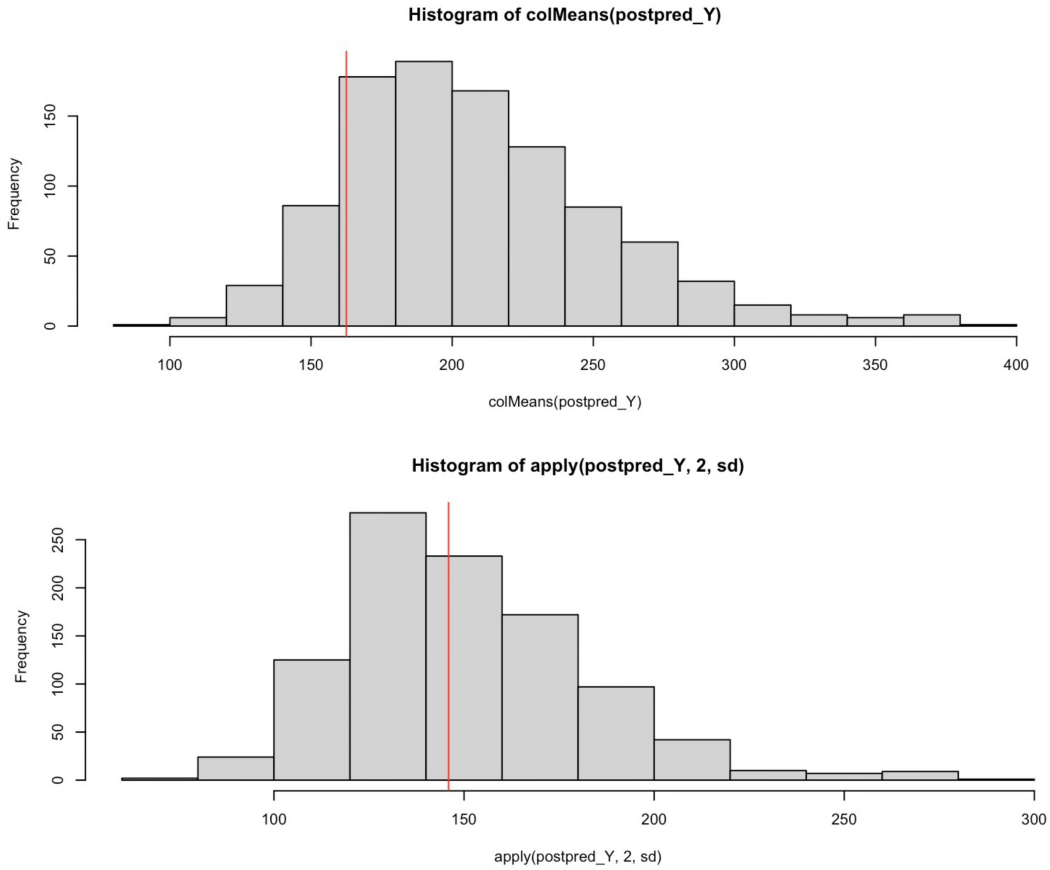


Figure A.12: histogram for CAR LEROUX big

Appendix B

Tables comparing DIC and WAIC

	single-year bym CAR model	single-year leroux CAR model
DIC	260.21164	259.55874
WAIC	254.89274	256.60850

Table B.1: DIC and WAIC for single-year bym and leroux CAR models

	Multiple-year bym CAR model	Multiple-year leroux CAR model
DIC	2639.8104	2598.6298
WAIC	2632.1961	2579.2742

Table B.2: DIC and WAIC for multiple-year bym and leroux CAR models

Model	MSE	RMSE	RMSPE
Bym (single-year)	488.1674	22.0945	0
Leroux (single-year)	493.9211	22.2243	0
Bym (multi-years)	655.8887	25.6103	0
Leroux (multi-years)	649.5979	25.4872	0
Poisson	1311.3722	36.2129	0

Table B.3: MSE, RMSE, and RMSPE Values for Various Models

Covariate	Bym single-year	Leroux single-year
Smoking Rate	0.0266 (0.0044, 0.0486)	0.0265 (0.0067, 0.0460)
Manufacturing	0.0073 (-0.0301, 0.0457)	0.0070 (-0.0275, 0.0422)
Construction	-0.0120 (-0.0637, 0.0392)	-0.0117 (-0.0596, 0.0342)
Emissions	-0.0002 (-0.0129, 0.0129)	-0.0005 (-0.0123, 0.0111)
Earnings	0.0007 (-0.0237, 0.0254)	-0.0005 (-0.0228, 0.0224)

Table B.4: Comparison of 95% interval estimates for covariates in the single-year CAR models

Covariate	Bym multi-years	Leroux multi-years
Smoking Rate	0.0132 (0.0082, 0.0183)	0.0127 (0.0079, 0.0175)
Manufacturing	-0.0056 (-0.0169, 0.0058)	-0.0048 (-0.0162, 0.0063)
Construction	-0.0119 (-0.0296, 0.0059)	-0.0118 (-0.0286, 0.0051)
Emissions	0.0078 (0.0038, 0.0102)	0.0068 (0.0038, 0.0098)
Earnings	-0.0198 (-0.0277, -0.0124)	-0.0198 (-0.0267, -0.0125)

Table B.5: Comparison of 95% interval estimates for covariates in the multi-year CAR models

Parameter	Bym single-year	Bym multi-years
τ^2	0.0165 (0.0040, 0.0424)	0.0245 (0.0166, 0.0340)
σ^2	0.0049 (0.0016, 0.0124)	0.0025 (0.0010, 0.0046)

Table B.6: Comparison of τ^2 and σ^2 for single-year and multi-years Bym models

Parameter	Leroux single-year	Leroux multi-years
τ^2	0.0178 (0.0063, 0.0392)	0.0288 (0.0217, 0.0374)
ρ	0.6143 (0.1158, 0.9623)	0.8522 (0.7168, 0.9449)

Table B.7: Comparison of τ^2 and ρ for single-year and multi-years Leroux models

Appendix C

Code Appendix

Contained herein is the comprehensive blueprint of the computational analysis that buttresses our research. This appendix meticulously catalogues each procedural phase, ensuring full disclosure and replicability of the research outcomes.

For the first part of the code appendix, we clean the dataset based on different characteristics of the covariates and lung cancer datasets. Then we use `mice` package and the `impute` function to implement the multiple imputation for the missing smoking data from 2008-2011. We then apply Rubin's Rule to aggregate the 40 imputation result to obtain the estimate of the Poisson model.

Our explorations delve into the realms of sensitivity analysis, wherein we fine-tune the imputation parameters and examine the stability of regression coefficients. This is complemented by bootstrap analysis, granting us insights into the variability and confidence intervals surrounding our model estimations

A novel incorporation in our methodological arsenal is the application of Moran's I statistic. This inclusion is pivotal, as it enables us to dissect the spatial correlation within the data, providing an essential correction for spatial dependency that could skew the results of lung cancer incidence analysis. We also use the `tmap` function to draw the paramter plot as well as the cluster plot.

The analytical narrative concludes with Bayesian CAR models. This framework is adeptly chosen for its fusion of prior knowledge with observed data, delivering a robust probabilistic approach to our predictions. The intricate MCMC simulations exemplify the depth of our computational explorations. Finally, we compute the rmse and rmspe for both the multi-year and single year models.

The specific codes are attached below.

Code Appendix

2024-02-25

```
# Load Required Libraries
library(dplyr)
library(forecast)
library(tidyr)
library(knitr)
library(stats)
library(ggplot2)
library(boot)
library(mice)
library(MASS)
require(JointAI)
library(brms)
library(loo)
library(sf)
library(spdep)
library(sp)
library(sf)
require(sf)
library(tmap)
library(CARBayes)
library(coda)


# Load the dataset
lung_cancer_data <- read.csv("Lung Cancer.csv")
smoking_data <- read.csv("Smoking.csv")
manuf_data <- read.csv("Manufacturing.csv")
constr_data <-  read.csv("Construction.csv")
emi_data <- read.csv("CO2.csv")
earning_data <- read.csv("Earnings.csv")
pop_data <- read.csv("Population.csv")




# Try Imputation Method using backwardcast by ARIMA (Times Series)
smoking_data_t <- t(smoking_data[-1])
smoking_ts <- ts(smoking_data_t, start = 2012, end = 2019, frequency = 1)
backcasts <- matrix(NA, nrow = 4, ncol = ncol(smoking_ts))
rownames(backcasts) <- 2008:2011

for(i in 1:ncol(smoking_ts)) {
  # Reverse the time series data for backcasting
  ts_data_reversed <- rev(smoking_ts[, i])
```

```r
  # Adjust the start time for the reversed series
  start_year <- 2019  # Assuming your data goes up to 2019
  ts_data_reversed <- ts(ts_data_reversed, start = start_year, frequency = 1)

  # Fit an ARIMA model to the reversed data
  fit <- auto.arima(ts_data_reversed)

  # Forecast (which is effectively backcasting) the missing years
  bc <- forecast(fit, h = 4)

  # Store the backcasted values in reverse order to align with original time
  backcasts[, i] <- rev(bc$mean)
}

# Convert backcasts to a data frame for easier viewing/manipulation
backcast_df <- as.data.frame(backcasts)
# Initialize an empty data frame to store backcasted values
backcast_df <- data.frame(
  council_area = rep(smoking_data$Council.Areas, each = 4),
  year = rep(2008:2011, times = ncol(smoking_ts)),
  backcasted_value = as.vector(backcasts)
)

colnames(backcast_df) <- c('council_area','year','backcasted_value')
names(backcast_df) <- c('Council.Areas','Year','Smoking_Rate')
backcast_df$Year <- paste0("X", backcast_df$Year)

# Standardise the Earning Data
earning_data[, -1] <- earning_data[, -1] / 1000

# Remove the last 15 rows
emi_data_cleaned <- emi_data %>%
  slice(1:(n() - 15)) %>%
  filter(Year >= 2008 & Year <= 2017)

emi_data_selected <- emi_data_cleaned %>%
  dplyr::select(Council.Areas, Year, Emissions_per_km2 )

# Pivot lung_cancer_data
Smoking_long <- pivot_longer(smoking_data,
                             cols = -Council.Areas,
                             names_to = "Year",
                             values_to = "Smoking_Rate")
# We only need the data from 2008-2017
Smoking_1 <- Smoking_long  %>%
  filter(Year >= 'X2008' & Year <= 'X2017')
merged_df <- merge(Smoking_1, backcast_df, by = c("Council.Areas",
                                                  "Year",
                                                  'Smoking_Rate'),
                                                      all = TRUE)


# Pivot lung_cancer_data
```

```r
lung_long <- pivot_longer(lung_cancer_data,
                          cols = -Council.Areas,
                          names_to = "Year",
                          values_to = "Lung_cancer_count")

# Pivot manuf_data
manuf_long <- pivot_longer(manuf_data,
                           cols = -Council.Areas,
                           names_to = "Year",
                           values_to = "Manufacturing_Employment_count")



# Pivot constr_data
constr_data_long <- pivot_longer(constr_data,
                                 cols = -Council.Areas,
                                 names_to = "Year",
                                 values_to = "Construction_Employment_count")




# Pivot earning_data
earning_data_long <- pivot_longer(earning_data,
                                  cols = -Council.Areas,
                                  names_to = "Year",
                                  values_to = "Earnings")
# Pivot pop_data
pop <- pivot_longer(pop_data,
                    cols = -Council.Areas,
                    names_to = "Year",
                    values_to = "Pop_count")

pop_data <- pop %>%
  filter(Year >= 'X2008' & Year <= 'X2017')


# Merge the data frames by Council.Areas and Year
merged_data <- merge(constr_data_long, pop_data, by = c("Council.Areas", "Year"))
# Perform the rates calculation on the merged data frame
merged_data$Construction_Rate <- merged_data$Construction_Employment_count *
  1000 / merged_data$Pop_count *100


merged_data_1 <- merge(manuf_long, pop_data, by = c("Council.Areas", "Year"))
merged_data_1$Manuf_Rate <- merged_data_1$Manufacturing_Employment_count *
  1000 / merged_data$Pop_count *100


# Final Merged Data
merged_reg_data <- cbind(lung_long[,1:3], Construction = merged_data[,5],
                         Manufacturing = merged_data_1[,5],
                         Emissions = emi_data_selected[,3],
```

```
                          Earnings = earning_data_long[,3],
                          Smoking_Rate = merged_df[,3],
                          Pop = pop_data[,3])
# Assuming your dataframe is named 'df' and the year column is named 'Year'
merged_reg_data$Smoking_Rate[merged_reg_data$Year %in% c("X2008",
                                                         "X2009",
                                                         "X2010",
                                                         "X2011")] <- NA

data <- merged_reg_data[, !colnames(merged_reg_data) %in% c('Council.Areas',
                                                            'Year')]

md.pattern(merged_reg_data)
# Perform MICE imputation
imputed_data <- mice(merged_reg_data, m=40, method='norm', maxit=10, seed=500)
plot(imputed_data)
stripplot(imputed_data, data = Smoking_Rate ~ .imp)
# Randomly choose the imputation data for the bayesian analysis
completed_data <- complete(imputed_data, action = 1)
```

```
# Fit Poisson and Negative Binomial Models on Imputed Datasets
Poi_model <- lapply(1:40, function(i) {
  completed_data <- complete(imputed_data, action = i)
  glm(Lung_cancer_count ~ Manufacturing + Construction + Emissions
      + Earnings + Smoking_Rate,
      data = completed_data, family = "poisson",
      offset = log(completed_data$Pop))
})


### Pool Model Results According to Rubin's Rules
pooled_results <- pool(Poi_model)
summary(pooled_results)


### Calculate and Print Mean AIC for Each Set of Models
aic_model <- sapply(Poi_model, AIC)
mean_aic_model <- mean(aic_model)
print(paste("Mean AIC for Poisson models with offset:", mean_aic_model))
```

```
### Sensitivity Analysis
# Sensitivity analysis by varying imputation parameters
imputed_data_varied <- mice(data, m=20, method='norm', maxit=5, seed=501)
# Comparing regression coefficients across varied imputation parameters
models_varied <- list()
for (i in 1:20) {
  completed_data_varied <- complete(imputed_data_varied, action = i)
  models_varied[[i]] <- glm(Lung_cancer_count ~ Manufacturing + Construction
                            + Emissions + Earnings + Smoking_Rate,
                            data = completed_data_varied, family = "poisson",
                            offset=log(Pop_count))
}
pooled_results_varied <- pool(models_varied)
summary(pooled_results_varied)
```

```r
### Bootstrap Analysis on Imputed Dataset
# Fitting Poisson model to bootstrap samples
fit_model <- function(data, indices) {
  boot_data <- data[indices, ]
  model <- glm(Lung_cancer_count ~ Manufacturing + Construction + Emissions
               + Earnings + Smoking_Rate,
               data = boot_data, family = "poisson", offset=log(Pop_count))
  coef(model)
}



num_coefficients <- 6
bootstrap_results <- matrix(NA, nrow = 40, ncol = num_coefficients)


# Iterates over each imputed dataset to perform bootstrap analysis.
for (i in 1:40) {
  # Retrieves the i-th imputed dataset.
  completed_data <- complete(imputed_data, action = i)

  # Performs bootstrap analysis on the imputed dataset.
  boot_res <- boot(data = completed_data, statistic = fit_model, R = 1000)

  # Calculates the mean of coefficients across bootstrap samples
  bootstrap_results[i, ] <- apply(boot_res$t, 2, mean)
}


# Computes the overall mean of the bootstrap results to estimate the variability
overall_coef_sd <- colMeans(bootstrap_results)

# Outputs the overall variability (standard deviation) of each coefficient.
print(overall_coef_sd)

# Calculates the 95% confidence intervals for the coefficients

ci_95 <- apply(bootstrap_results,
               2, function(x) quantile(x, probs = c(0.025, 0.975)))

# Converts the confidence intervals into a more readable data frame format.
ci_95_df <- as.data.frame(t(ci_95))

# Assigns names to the columns
colnames(ci_95_df) <- c("2.5%", "97.5%")
# Assigns names to the rows based on the predictors and intercept.
rownames(ci_95_df) <- c("Intercept", "Manufacturing", "Construction",
                        "Emissions", "Earnings", "Smoking_Rate")

# Displays the confidence intervals for each coefficient.
print(ci_95_df)
```

```r
# WAIC
# Convert the model to a Bayesian framework using brms
bayesian_model <- brm(
  formula = Lung_cancer_count ~ Manufacturing + Construction + Emissions
  + Earnings + Smoking_Rate + offset(log(Pop_count)),
  family = poisson(),
  data = completed_data,
  prior <- c(
    set_prior("normal(0, 10)", class = "b"),
    set_prior("normal(0, 10)", class = "Intercept")
  ),
  chains = 4,
  iter = 2000,
  warmup = 1000,
  cores = 4,
  seed = 123,
  save_pars = save_pars(all = TRUE)  # Ensure all parameters are saved
)


# Calculate WAIC with moment matching
waic_result <- loo(bayesian_model, save_psis = TRUE)


# Print WAIC result
print(waic_result)
WAIC_Poi <- waic_result$estimates["looic", "Estimate"]
```

```r
### Spatial Analysis Section
# Ratio Data
shape <- read_sf(dsn = "/home/baoqizhang/Desktop/Projects/Shape")
shape <- st_read("/home/baoqizhang/Desktop/Projects/Shape")
# Unified council areas name for shape file and the lung cancer data
shape$local_auth <- gsub("Eilean Siar", "Na h-Eileanan Siar", shape$local_auth)
lung_cancer_data <- read.csv("Lung Cancer.csv")
pop_data <- read.csv("Population.csv")

calculate_ratio <- function(lung_cancer_data, pop_data) {
  # Extract the year columns from both data frames
  years <- colnames(lung_cancer_data)[-1]

  # Initialize an empty data frame to store the results
  result <- data.frame(council.areas = pop_data$Council.Areas)

  # Loop through each year and calculate the ratio
  for (year in years) {
    lung_cancer_col <- as.numeric(lung_cancer_data[, year])
    pop_col <- as.numeric(pop_data[, year])

    # Check for missing or non-numeric values
    if (any(is.na(lung_cancer_col)) || any(is.na(pop_col))) {
      warning(paste("Skipping year", year,
                    "due to missing or non-numeric values"))
      result[[year]] <- NA
```

```
    } else {
      # Calculate the ratio (lung cancer cases / population)
      ratio <- (lung_cancer_col / pop_col)*100
      result[[year]] <- ratio
    }
  }

  return(result)
}

ratio_data <- calculate_ratio(lung_cancer_data, pop_data)
# Calculate the average ratio across specified columns, ignoring NA values
ratio_data$mean_ratio = rowMeans(ratio_data[, 2:ncol(ratio_data)], na.rm = TRUE)
# Display the first few rows of the updated ratio dataset for verification
head(ratio_data)

lung_cancer_data <- read.csv("Lung Cancer.csv")
lung_cancer_data$mean_count = rowMeans(lung_cancer_data[, 2:ncol(lung_cancer_data)])
# Display the first few rows of the lung cancer data to ensure accuracy
head(lung_cancer_data)

# Merge ratio and lung cancer datasets based on council areas
comb <- merge(ratio_data, lung_cancer_data, by.x = "council.areas",
              by.y = "Council.Areas")
# Further merge the combined dataset with spatial boundary data
merged_data_1 <- merge(shape, comb, by.x = "local_auth", by.y = "council.areas")


# Converting shape file to the Spatial File
shape_sp <- as(merged_data_1, "Spatial")
```

```
# Identification and Analysis of Spatial Neighbors

# Conversion of the Spatial*DataFrame to an 'sf' object
shape_sf <- st_as_sf(shape_sp)

# Calculation of centroids for each spatial feature
centroids <- st_centroid(shape_sf)

# Extraction of x and y coordinates from the calculated centroids
centroids_coords <- st_coordinates(centroids)

# Inclusion of centroid coordinates back into the 'sf' object
# This enriches the spatial dataset with precise location data for each area
shape_sf$x <- centroids_coords[, 1]
shape_sf$y <- centroids_coords[, 2]

# Identification of spatial neighbors based on the Queen contiguity criterion
# This method determines adjacency based on shared borders or vertices
nb_q_sp <- poly2nb(shape_sf)

# Manual adjustments for specific spatial relationships
nb_q_sp[[20]] <- c(nb_q_sp[[20]], 4)
nb_q_sp[[4]] <- c(nb_q_sp[[4]], 20)
```

```r
nb_q_sp[[27]] <- c(nb_q_sp[[27]], 1)
nb_q_sp[[1]] <- c(nb_q_sp[[1]], 27)
nb_q_sp[[27]] <- c(nb_q_sp[[27]], 23)
nb_q_sp[[23]] <- c(nb_q_sp[[23]], 27)
nb_q_sp[[1]] <- c(nb_q_sp[[1]], 23)
nb_q_sp[[23]] <- c(nb_q_sp[[23]], 1)
nb_q_sp[[16]] <- c(nb_q_sp[[16]], 23)
nb_q_sp[[23]] <- c(nb_q_sp[[23]], 16)
nb_q_sp[[16]] <- c(nb_q_sp[[16]], 20)
nb_q_sp[[20]] <- c(nb_q_sp[[20]], 16)

# Refinement of the neighbors list to exclude non-neighbors
nb_q_sp <- lapply(nb_q_sp, function(neighbors) neighbors[neighbors != 0])
# Summary of the neighbors list to validate the adjustments
summary(nb_q_sp)

# Construction of an adjacency matrix from the neighbors list
# The adjacency matrix represents the neighbor relationships between areas
n_neigh <- nrow(shape_sp)
adj_mat <- matrix(data = 0, nrow = n_neigh, ncol = n_neigh)
for(i in 1:n_neigh) {
  adj_mat[i, nb_q_sp[[i]]] <- 1
}

# Conversion of the neighbors list into a 'nb' class object
nb_q_sp <- lapply(nb_q_sp, as.integer)
class(nb_q_sp) <- "nb"
attr(nb_q_sp, "region.id") <- as.character(1:32)

# Creation of a 'listw' object with a zero policy
lw_q_B <- nb2listw(nb_q_sp, style="B", zero.policy=TRUE)

# Perform Shapiro-Wilk normality test
shapiro_result <- shapiro.test(shape_sf$mean_ratio)

# Print the results
print(shapiro_result)

# Execution of the Global Moran's I Test using the mean ratio data
moran_result <- moran.test(shape_sp$mean_ratio, lw_q_B)
# Presentation of the Moran's I Test results
print(moran_result)
# Moran's I scatterplot to illustrate the spatial autocorrelation
moran.plot(shape_sp$mean_ratio, lw_q_B, return_df = TRUE)
```

```r
### Visualization of Neighborhood Structures
# Calculate the number of neighbors for each spatial unit
num_neighbors <- sapply(nb_q_sp, length)

# Generate a color palette to visually differentiate spatial units
num_breaks <- length(unique(num_neighbors))
colors <- colorRampPalette(c("white", "cornflowerblue"))(num_breaks)

# Map each unique neighbor count to a specific color index
```

```r
color_mapping <- setNames(seq_len(num_breaks), sort(unique(num_neighbors)))

# Assign colors to each spatial unit according to its number of neighbors
area_colors <- colors[color_mapping[as.character(num_neighbors)]]

# Plot the spatial framework
plot(st_geometry(shape_sf), col = area_colors, border = "black")

# Create legend labels
legend_labels <- paste(names(color_mapping), "neighbors")
legend_colors <- colors

# Display a legend on the plot to interpret the color scheme
legend("topright", legend = legend_labels, fill = legend_colors, cex = 0.7)

# Illustrate the neighborhood relationships by connecting them
for (i in 1:length(nb_q_sp)) {
  # Iterate through each spatial unit to identify its neighbors
  neighbors <- nb_q_sp[[i]]
  for (j in neighbors) {
    # Draw lines connecting each unit to its neighbors, emphasizing the
    # neighborhood structure
    lines(rbind(centroids_coords[i,], centroids_coords[j,]), col="black")
  }
}

# Define a semi-transparent color for highlighting centroids
transparent_red <- rgb(1, 0, 0, alpha = 0.5)

# Centroids represent the geometric centers of spatial units,
# providing a focal point for visualizing neighborhood connections
points(centroids_coords[,1], centroids_coords[,2], pch = 21, col = "black",
       bg = transparent_red, cex = 1.5)
```

```r
### Local Moran's I
# Calculate Local Moran's I
local_moran <- localmoran(shape_sf$mean_ratio, lw_q_B)
print(local_moran)


tmap_mode("plot")
shape_sp_1 <- st_as_sf(shape_sp)

# Add the Local Moran's I statistics to the sf object
shape_sp_1$lmI <- local_moran[, "Ii"] # local Moran's I
shape_sp_1$lmZ <- local_moran[, "Z.Ii"] # z-scores
shape_sp_1$lmp <- local_moran[, "Pr(z != E(Ii))"] # p-values

# Map 1: Lung Cancer Mean Ratio
p1 <- tm_shape(shape_sp_1) +
  tm_polygons(col = "mean_ratio", title = "Lung Cancer Rate",
              style = "quantile") +
  tm_layout(legend.outside = TRUE,
            legend.text.size = 0.9)
```

```r
p1
# Map 2: Local Moran's I
p2 <- tm_shape(shape_sp_1) +
  tm_polygons(col = "lmI", title = "Local Moran's I",
              style = "quantile") +
  tm_layout(legend.outside = TRUE,
            legend.text.size = 0.9)
p2
# Map 3: Z-score
p3 <- tm_shape(shape_sp_1) +
  tm_polygons(col = "lmZ", title = "Z-score",
              # Use qnorm(0.975) for two-sided test
              breaks = c(-Inf, qnorm(0.95), Inf),
               # Red for significant positive autocorrelation
              palette = c("red", "white")) +
  tm_layout(legend.outside = TRUE,
            legend.text.size = 0.9)
p3
# Map 4: p-value
p4 <- tm_shape(shape_sp_1) +
  tm_polygons(col = "lmp", title = "p-value",
              # Use 0.025 and 0.975 for two-sided test
              breaks = c(-Inf, 0.05, Inf),
              # Red for p-value less than 0.05
              palette = c("red", "white")) +
  tm_layout(legend.outside = TRUE,
            legend.text.size = 0.9)
p4
```

```r
# LISA Cluster Plot
mp <- moran.plot(as.vector(scale(shape_sp_1$mean_ratio)), lw_q_B)
shape_sp_1$lmp <- local_moran[, 5]

shape_sp_1$quadrant <- NA
# high-high
shape_sp_1[(mp$x >= 0 & mp$wx >= 0) & (shape_sp_1$lmp <= 0.05), "quadrant"]<- 1
# low-low
shape_sp_1[(mp$x <= 0 & mp$wx <= 0) & (shape_sp_1$lmp <= 0.05), "quadrant"]<- 2
# high-low
shape_sp_1[(mp$x >= 0 & mp$wx <= 0) & (shape_sp_1$lmp <= 0.05), "quadrant"]<- 3
# low-high
shape_sp_1[(mp$x <= 0 & mp$wx >= 0) & (shape_sp_1$lmp <= 0.05), "quadrant"]<- 4
# non-significant
shape_sp_1[(shape_sp_1$lmp > 0.05), "quadrant"] <- 5
tm_shape(shape_sp_1) + tm_fill(col = "quadrant", title = "",
                         breaks = c(1, 2, 3, 4, 5, 6),
                         palette =  c("red", "blue", "lightpink", "skyblue2",
                                      "white"),
                         labels = c("High-High", "Low-Low", "High-Low",
                                    "Low-High", "Non-significant")) +
  tm_legend(text.size = 1)  + tm_borders(alpha = 0.5) +
  tm_layout(frame = FALSE,  title = "Clusters")  +
  tm_layout(legend.outside = TRUE)
```

```r
# Define the necessary variables
weights <- nb2listw(nb_q_sp, style="B", zero.policy=TRUE)
neighbors_list <- as(weights, "listw")$neighbours
W_matrix <- nb2mat(nb_q_sp, style="B", zero.policy=TRUE)

# MCMC parameters
M.burnin <- 10000       # Number of burn-in iterations (discarded)
M <- 100000               # Number of iterations retained

# use 2016 model to predict 2017

# car model for 2016 since need to predict for 2017
combined_data_2016 <- filter(completed_data, Year == "X2016")


# BYM
set.seed(444)           # For reproducability
MCMC_bym_all <- S.CARbym(
  formula = Lung_cancer_count ~ offset(log(Pop_count)) + Smoking_Rate
  + Manufacturing +
    Construction + Emissions + Earnings,
  data = combined_data_2016,
  family = "poisson",
  W = W_matrix,
  burnin = M.burnin,
  n.sample = M.burnin + M,    # Total iterations
  n.chains = 3,
  n.cores = 3,
  verbose = FALSE
)

# Broad summary of results
print(MCMC_bym_all$summary.results)

# Additional results on model fit criteria
MCMC_bym_all$modelfit

# model check
beta_bym_all <- MCMC_bym_all$samples$beta
gelman_results_bym_all <- gelman.diag(beta_bym_all)
print(gelman_results_bym_all) #1 converges

plot(beta_bym_all)
autocorr.plot(beta_bym_all)
gelman.plot(beta_bym_all)

# Check posterior predict model
psi <- MCMC_bym_all$samples$psi

samples_beta_bym_all <- do.call(rbind, lapply(MCMC_bym_all$samples$beta,
                                              function(x) x))
samples_psi_bym_all <- do.call(rbind, lapply(MCMC_bym_all$samples$psi,
                                             function(x) x))
```

```r
Nsamples = 1000
postpred_Y <- array(NA,dim=c(32,Nsamples))
postpred_mean <- array(NA,dim=c(32,Nsamples))

for (i_s in 1:Nsamples) {
  # randomly chose a set of samples after the combining samples
  intercept_s <- samples_beta_bym_all[s, 1]
  beta_smoking_s <- samples_beta_bym_all[s, 2]
  beta_construction_s <- samples_beta_bym_all[s, 3]
  beta_earnings_s <- samples_beta_bym_all[s, 4]
  beta_manufacturing_s <- samples_beta_bym_all[s, 5]
  beta_emissions_s <- samples_beta_bym_all[s,6]
  psi_means_s <- samples_psi_bym_all[s, ]

  # calculate the average for the prediction
  postpred_mean[, i_s] <- with(combined_data_2016,
                               exp(intercept_s +
                                     beta_smoking_s * Smoking_Rate +
                                     beta_construction_s * Construction +
                                     beta_manufacturing_s * Manufacturing +
                                     beta_earnings_s * Earnings +
                                     beta_emissions_s * Emissions + psi_means_s +
                                     offset(log(Pop_count))
                               ))
  # generating prediction counts
  postpred_Y[, i_s] <- rpois(32, lambda = postpred_mean[, i_s])
}

# Compare actual and prediction
hist(colMeans(postpred_Y))
abline(v = mean(combined_data_2016$Lung_cancer_count), col = "red")

hist(apply(postpred_Y, 2, sd))
abline(v = sd(combined_data_2016$Lung_cancer_count), col = "red")

# predict
car_2017 <- filter(completed_data, Year == "X2017")

# Extract posterior means
intercept_bym <- MCMC_bym_all$summary.results["(Intercept)", "Mean"]
beta_smoking_bym <- MCMC_bym_all$summary.results["Smoking_Rate", "Mean"]
beta_construction_bym <- MCMC_bym_all$summary.results["Construction", "Mean"]
beta_earnings_bym <- MCMC_bym_all$summary.results["Earnings", "Mean"]
beta_manufacturing_bym <- MCMC_bym_all$summary.results["Manufacturing", "Mean"]
beta_emissions_bym <- MCMC_bym_all$summary.results["Emissions", "Mean"]
psi_means_bym = colMeans(samples_psi_bym_all)

# Predicting lung cancer cases for 2017
car_2017$predicted_lung_cancer_bym <- with(car_2017,
                                           exp(intercept_bym +
                                 beta_smoking_bym * Smoking_Rate +
                                 beta_construction_bym * Construction +
                                 beta_manufacturing_bym * Manufacturing +
```

```
                                beta_earnings_bym * Earnings +
                                beta_emissions_bym * Emissions + psi_means_bym +
                                        offset(log(Pop_count))
                                    )
)


# Add a column to compare the predicted cases to actual cases
car_2017$comparison_b <- with(car_2017, predicted_lung_cancer_bym - Lung_cancer_count)
mean(car_2017$comparison_b^2) # 489.4742 BYM(1 year)

# Create a scatter plot
ggplot(car_2017, aes(x = Lung_cancer_count, y = predicted_lung_cancer_bym)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
  labs(x = "Actual Lung Cancer Count", y = "Predicted Lung Cancer Count",
       title = "Actual vs Predicted Lung Cancer Cases") +
  theme_minimal()
```

```
# LEROUX
set.seed(444)
MCMC_leroux_all <- S.CARleroux(
  formula = Lung_cancer_count ~ offset(log(Pop_count)) + Smoking_Rate
  + Manufacturing + Construction + Emissions + Earnings,
  data = combined_data_2016,
  family = "poisson",
  W = W_matrix,
  burnin = M.burnin,
  n.sample = M.burnin + M,     # Total iterations
  n.chains = 3,
  n.cores = 3,
  verbose = FALSE
)

# Broad summary of results
print(MCMC_leroux_all$summary.results)

# Additional results on model fit criteria
MCMC_leroux_all$modelfit

# model check
beta_le_all <- MCMC_leroux_all$samples$beta
gelman_results_le_all <- gelman.diag(beta_le_all)
print(gelman_results_le_all) #1 converge

plot(beta_le_all)
autocorr.plot(beta_le_all)
gelman.plot(beta_le_all)

# Check posterior predict model
phi <- MCMC_leroux_all$samples$phi

samples_beta_le_all <- do.call(rbind, lapply(MCMC_leroux_all$samples$beta,
                                             function(x) x))
```

```r
samples_phi_le_all <- do.call(rbind, lapply(MCMC_leroux_all$samples$phi,
                                            function(x) x))

Nsamples = 1000
postpred_Y <- array(NA,dim=c(32,Nsamples))
postpred_mean <- array(NA,dim=c(32,Nsamples))

for (i_s in 1:Nsamples) {
  # Choose randomly a set of samples
  s = 10*i_s
  intercept_s_le <- samples_beta_le_all[s, 1]
  beta_smoking_s_le <- samples_beta_le_all[s, 2]
  beta_construction_s_le <- samples_beta_le_all[s, 3]
  beta_earnings_s_le <- samples_beta_le_all[s, 4]
  beta_manufacturing_s_le <- samples_beta_le_all[s, 5]
  beta_emissions_s_le <- samples_beta_le_all[s, 6]
  phi_means_s_le <- samples_phi_le_all[s, ]

  # Calculate the prediction average
  postpred_mean[, i_s] <- with(combined_data_2016,
                               exp(intercept_s_le +
                                     beta_smoking_s_le * Smoking_Rate +
                                     beta_construction_s_le * Construction +
                                     beta_manufacturing_s_le * Manufacturing +
                                     beta_earnings_s_le * Earnings +
                                     beta_emissions_s_le * Emissions +
                                     phi_means_s_le +
                                     offset(log(Pop_count))
                               ))
  # Generating prediction counts
  postpred_Y[, i_s] <- rpois(32, lambda = postpred_mean[, i_s])
}

# Compare actual and prediction
hist(colMeans(postpred_Y))
abline(v = mean(combined_data_2016$Lung_cancer_count), col = "red")

hist(apply(postpred_Y, 2, sd))
abline(v = sd(combined_data_2016$Lung_cancer_count), col = "red")

# predict for 2017
# Extract posterior means
intercept_le <- MCMC_leroux_all$summary.results["(Intercept)", "Mean"]
beta_smoking_le <- MCMC_leroux_all$summary.results["Smoking_Rate", "Mean"]
beta_construction_le <- MCMC_leroux_all$summary.results["Construction", "Mean"]
beta_earnings_le <- MCMC_leroux_all$summary.results["Earnings", "Mean"]
beta_manufacturing_le <- MCMC_leroux_all$summary.results["Manufacturing", "Mean"]
beta_emissions_le <- MCMC_leroux_all$summary.results["Emissions", "Mean"]
phi_means_le <- colMeans(samples_phi_le_all)

# Predicting lung cancer cases for 2017
car_2017$predicted_lung_cancer_le <- with(car_2017,
                                          exp(intercept_le +
```

```
                                          beta_smoking_le * Smoking_Rate +
                                        beta_construction_le * Construction +
                                        beta_manufacturing_le * Manufacturing +
                                        beta_earnings_le * Earnings +
                                        beta_emissions_le * Emissions +
                                        phi_means_le +
                                        offset(log(Pop_count))))


# Add a column to compare the predicted cases to actual cases
car_2017$comparison_le <- with(car_2017, predicted_lung_cancer_le - Lung_cancer_count)
mean(car_2017$comparison_le^2) # 498.3557 Leroux(1 year)

# Create a scatter plot
ggplot(car_2017, aes(x = Lung_cancer_count, y = predicted_lung_cancer_le)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
  labs(x = "Actual Lung Cancer Count", y = "Predicted Lung Cancer Count",
       title = "Actual vs Predicted Lung Cancer Cases") +
  theme_minimal()
```

```
# 9 Year predict 2017
W_big_2 <- matrix(0, nrow = 288, ncol = 288)

# Replace diagonal with W_matrix
for (i in 0:8) {
  rows <- (1:32) + i * 32
  cols <- (1:32) + i * 32
  W_big_2[rows, cols] <- W_matrix
}

library(dplyr)
filtered_data <- completed_data %>%
  filter(!grepl("X2017", Year))
completed_data_car <-completed_data[order(completed_data$Year), ]
filtered_data_car <- completed_data_car[completed_data_car$Year != "X2017", ]

M.burnin <- 10000       # Number of burn-in iterations (discarded)
M <- 100000               # Number of iterations retained



# BYM
set.seed(444)          # For reproducability
MCMC_big_bym_all <- S.CARbym(
  formula = Lung_cancer_count ~ offset(log(Pop_count)) + Smoking_Rate
  + Manufacturing +
    Construction + Emissions + Earnings,
  data = filtered_data_car,
  family = "poisson",
  W = W_big_2,
  burnin = M.burnin,
  n.sample = M.burnin + M,     # Total iterations
  n.chains = 3,
```

```
  n.cores = 3,
  verbose = FALSE
)

# Broad summary of results
print(MCMC_big_bym_all$summary.results)

# Additional results on model fit criteria
MCMC_big_bym_all$modelfit

# model check
beta_bym_all_big <- MCMC_big_bym_all$samples$beta
gelman_results_bym_all_big <- gelman.diag(beta_bym_all_big)
print(gelman_results_bym_all_big) # 1.01

plot(beta_bym_all_big)
autocorr.plot(beta_bym_all_big)
gelman.plot(beta_bym_all_big)

# Check posterior predict model
psi <- MCMC_big_bym_all$samples$psi

samples_beta_bym_all_big <- do.call(rbind, lapply(MCMC_big_bym_all$samples$beta,
                                                  function(x) x))
samples_psi_bym_all_big <- do.call(rbind, lapply(MCMC_big_bym_all$samples$psi,
                                                 function(x) x))

Nsamples = 1000
postpred_Y <- array(NA,dim=c(288,Nsamples))
postpred_mean <- array(NA,dim=c(288,Nsamples))

for (i_s in 1:Nsamples) {
  # Randomly select a set of parameter samples
  s = 10*i_s
  intercept_s <- samples_beta_bym_all_big[s, 1]
  beta_smoking_s <- samples_beta_bym_all_big[s, 2]
  beta_construction_s <- samples_beta_bym_all_big[s, 3]
  beta_earnings_s <- samples_beta_bym_all_big[s, 4]
  beta_manufacturing_s <- samples_beta_bym_all_big[s, 5]
  beta_emissions_s <- samples_beta_bym_all_big[s,6]
  psi_means_s <- samples_psi_bym_all_big[s, ]

  # Calculate the mean value of the prediction
  postpred_mean[, i_s] <- with(filtered_data_car,
                               exp(intercept_s +
                                     beta_smoking_s * Smoking_Rate +
                                     beta_construction_s * Construction +
                                     beta_manufacturing_s * Manufacturing +
                                     beta_earnings_s * Earnings +
                                     beta_emissions_s * Emissions + psi_means_s +
                                     offset(log(Pop_count))
                               ))
  # Generate predictive counts
```

```r
  postpred_Y[, i_s] <- rpois(288, lambda = postpred_mean[, i_s])
}

# Comparison of predicted and actual observed data
hist(colMeans(postpred_Y))
abline(v = mean(filtered_data_car$Lung_cancer_count), col = "red")

hist(apply(postpred_Y, 2, sd))
abline(v = sd(filtered_data_car$Lung_cancer_count), col = "red")

# predict!
car_big_2017 <- filter(completed_data, Year == "X2017")

# Extract posterior means
intercept_big <- MCMC_big_bym_all$summary.results["(Intercept)", "Mean"]
beta_smoking_big <- MCMC_big_bym_all$summary.results["Smoking_Rate", "Mean"]
beta_construction_big <- MCMC_big_bym_all$summary.results["Construction", "Mean"]
beta_earnings_big <- MCMC_big_bym_all$summary.results["Earnings", "Mean"]
beta_manufacturing_big <- MCMC_big_bym_all$summary.results["Manufacturing", "Mean"]
beta_emissions_big <- MCMC_big_bym_all$summary.results["Emissions", "Mean"]

psi_all_chains <- do.call(rbind, MCMC_big_bym_all$samples$psi)
psi_means <- colMeans(psi_all_chains)
psi_means_big_chain <- matrix(psi_means, nrow = 32, ncol = 9, byrow = FALSE)
psi_row_averages <- rowMeans(psi_means_big_chain)

# Predicting lung cancer cases for 2017
car_big_2017$predicted_lung_cancer_bym_big <- with(car_big_2017,
                                                   exp(intercept_big +
                                                 beta_smoking_big * Smoking_Rate +
                                           beta_construction_big * Construction +
                                       beta_manufacturing_big * Manufacturing +
                                                 beta_earnings_big * Earnings +
                                               beta_emissions_big * Emissions +
                                                             psi_row_averages +
                                                     offset(log(Pop_count))))


# Add a column to compare the predicted cases to actual cases
car_big_2017$comparison_big_B <- with(car_big_2017,
                          predicted_lung_cancer_bym_big - Lung_cancer_count)
mean(car_big_2017$comparison_big_B^2) # 671.8069 BYM multi_year

# Create a scatter plot
ggplot(car_big_2017, aes(x = Lung_cancer_count,
                         y = predicted_lung_cancer_bym_big)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
  labs(x = "Actual Lung Cancer Count", y = "Predicted Lung Cancer Count",
       title = "Actual vs Predicted Lung Cancer Cases") +
  theme_minimal()
```

```r
# LEROUX
set.seed(444)          # For reproducability
```

```
MCMC_big_le_all <- S.CARleroux(
  formula = Lung_cancer_count ~ offset(log(Pop_count)) +
    Smoking_Rate + Manufacturing +
    Construction + Emissions + Earnings,
  data = filtered_data_car,
  family = "poisson",
  W = W_big_2,
  burnin = M.burnin,
  n.sample = M.burnin + M,     # Total iterations
  n.chains = 3,
  n.cores = 3,
  verbose = FALSE
)

# Broad summary of results
print(MCMC_big_le_all$summary.results)

# Additional results on model fit criteria
MCMC_big_le_all$modelfit



# Check posterior predict model
phi <- MCMC_big_le_all$samples$phi

samples_beta_le_all_big <- do.call(rbind,
                           lapply(MCMC_big_le_all$samples$beta, function(x) x))
samples_phi_bym_all_big <- do.call(rbind,
                           lapply(MCMC_big_le_all$samples$phi, function(x) x))

Nsamples = 1000
postpred_Y <- array(NA,dim=c(288,Nsamples))
postpred_mean <- array(NA,dim=c(288,Nsamples))

for (i_s in 1:Nsamples) {

  s = 10*i_s
  intercept_s <- samples_beta_le_all_big[s, 1]
  beta_smoking_s <- samples_beta_le_all_big[s, 2]
  beta_construction_s <- samples_beta_le_all_big[s, 3]
  beta_earnings_s <- samples_beta_le_all_big[s, 4]
  beta_manufacturing_s <- samples_beta_le_all_big[s, 5]
  beta_emissions_s <- samples_beta_le_all_big[s,6]
  phi_means_s <- samples_phi_bym_all_big[s, ]


  postpred_mean[, i_s] <- with(filtered_data_car,
                               exp(intercept_s +
                                     beta_smoking_s * Smoking_Rate +
                                     beta_construction_s * Construction +
                                     beta_manufacturing_s * Manufacturing +
                                     beta_earnings_s * Earnings +
                                     beta_emissions_s * Emissions + phi_means_s +
```

```
                                    offset(log(Pop_count))
                                ))

  postpred_Y[, i_s] <- rpois(288, lambda = postpred_mean[, i_s])
}




# Extract posterior means
intercept_big <- MCMC_big_le_all$summary.results["(Intercept)", "Mean"]
beta_smoking_big <- MCMC_big_le_all$summary.results["Smoking_Rate", "Mean"]
beta_construction_big <- MCMC_big_le_all$summary.results["Construction", "Mean"]
beta_earnings_big <- MCMC_big_le_all$summary.results["Earnings", "Mean"]
beta_manufacturing_big <- MCMC_big_le_all$summary.results["Manufacturing", "Mean"]
beta_emissions_big <- MCMC_big_le_all$summary.results["Emissions", "Mean"]

phi_all_chains <- do.call(rbind, MCMC_big_le_all$samples$phi)
phi_means <- colMeans(phi_all_chains)
phi_means_big_chain <- matrix(phi_means, nrow = 32, ncol = 9, byrow = FALSE)
phi_row_averages <- rowMeans(phi_means_big_chain)

# Predicting lung cancer cases for 2017
car_big_2017$predicted_lung_cancer_le_big <- with(car_big_2017,
                              exp(intercept_big +
                              beta_smoking_big * Smoking_Rate +
                              beta_construction_big * Construction +
                              beta_manufacturing_big * Manufacturing +
                              beta_earnings_big * Earnings +
                              beta_emissions_big * Emissions + phi_row_averages +
                              offset(log(Pop_count))
                                                  )
)


# Add a column to compare the predicted cases to actual cases
car_big_2017$comparison_big_L <- with(car_big_2017,
                              predicted_lung_cancer_le_big - Lung_cancer_count)
mean(car_big_2017$comparison_big_L^2) # 666.2226 Leroux(multi-year)



poisson_model_4 <- glm(Lung_cancer_count ~ Manufacturing + Construction
                       + Emissions + Earnings + Smoking_Rate,
                       offset=offset(log(Pop_count)),
                       data = completed_data,
                       family = "poisson")
summary(poisson_model_4)s


# prepare the model for 2008-2016 since we need to predict for 2017
filtered_data <- completed_data %>%
  filter(completed_data$Year >= "X2008" & completed_data$Year <= "X2016")
```

```r
# final model for glm in poisson
poisson_model_4 <- glm(Lung_cancer_count ~ Manufacturing + Construction
                       + Emissions + Earnings + Smoking_Rate,
                       offset=offset(log(Pop_count)),
                       data = filtered_data,
                       family = "poisson")
summary(poisson_model_4)
plot(poisson_model_4)

# predict for 2017
y_2017 <- completed_data %>% filter(Year == "X2017")

# Predicting lung cancer cases in 2017
predicted_counts <- predict(poisson_model_4, y_2017, type = "response")

# Add the predicted counts to your y_2017 dataset
y_2017$predicted_lung_cancer = predicted_counts

# Compute absolute value
y_2017$absolute_error <- abs(y_2017$Lung_cancer_count - y_2017$predicted_lung_cancer)

# View the results
View(y_2017)

# calculate the MSE
squared_errors <- (y_2017$Lung_cancer_count - y_2017$predicted_lung_cancer)^2
mse <- mean(squared_errors)
print(mse) ## 1311.372



# prepare the model for 2008-2016 since we need to predict for 2017
filtered_data <- completed_data %>%
  filter(completed_data$Year >= "X2008" & completed_data$Year <= "X2016")
```

```r
#make a 320x320 matrix filled with 0
W_big_3 <- matrix(0, nrow = 320, ncol = 320)

# Replace diagonal with W_matrix
for (i in 0:9) {
  rows <- (1:32) + i * 32
  cols <- (1:32) + i * 32
  W_big_3[rows, cols] <- W_matrix
}



# BYM plus all covariates
set.seed(444)          # For reproducability
MCMC_bym_allyear <- S.CARbym(
  formula = Lung_cancer_count ~ offset(log(Pop_count))
  + Smoking_Rate + Manufacturing +
    Construction + Emissions + Earnings,
```

```
  data = completed_data_car,
  family = "poisson",
  W = W_big_3,
  burnin = M.burnin,
  n.sample = M.burnin + M,     # Total iterations
  n.chains = 3,
  n.cores = 3,
  verbose = FALSE
)

# Broad summary of results
print(MCMC_bym_allyear$summary.results)

# Additional results on model fit criteria
MCMC_bym_allyear$modelfit
WAIC_BYM <- MCMC_bym_allyear$modelfit["WAIC"]

# LEROUX plus all covariates
set.seed(444)           # For reproducability
MCMC_leroux_allyear <- S.CARleroux(
  formula = Lung_cancer_count ~ offset(log(Pop_count)) +
    Smoking_Rate + Manufacturing +
    Construction + Emissions + Earnings,
  data = completed_data_car,
  family = "poisson",
  W = W_big_3,
  burnin = M.burnin,
  n.sample = M.burnin + M,     # Total iterations
  n.chains = 3,
  n.cores = 3,
  verbose = FALSE
)

# Broad summary of results
print(MCMC_leroux_allyear$summary.results)

# Additional results on model fit criteria
MCMC_leroux_allyear$modelfit
WAIC_LEROUX <- MCMC_leroux_allyear$modelfit["WAIC"]



# RMSE calculated
# bym 1
car_2017$comparison_b <- with(car_2017,
                              predicted_lung_cancer_bym - Lung_cancer_count)
mean(car_2017$comparison_b^2) # 489.4742

rmse1 <- mean((car_2017$comparison_b/car_2017$Lung_cancer_count)^2)
sqrt(rmse1)

# le 1
car_2017$comparison_le <- with(car_2017,
```

```
                                predicted_lung_cancer_le - Lung_cancer_count)
mean(car_2017$comparison_le^2) # 498.3557

rmse2 <- mean((car_2017$comparison_le/car_2017$Lung_cancer_count)^2)
sqrt(rmse2)

# bym multi
car_big_2017$comparison_big_B <- with(car_big_2017,
                              predicted_lung_cancer_bym_big - Lung_cancer_count)
mean(car_big_2017$comparison_big_B^2) # 671.8069

rmse3 <- mean((car_big_2017$comparison_big_B/car_big_2017$Lung_cancer_count)^2)
sqrt(rmse3)

# le multi
car_big_2017$comparison_big_L <- with(car_big_2017,
                              predicted_lung_cancer_le_big - Lung_cancer_count)
mean(car_big_2017$comparison_big_L^2) # 666.2226

rmse4 <- mean((car_big_2017$comparison_big_L/car_big_2017$Lung_cancer_count)^2)
sqrt(rmse4)

# poisson
squared_errors <- (y_2017$Lung_cancer_count - y_2017$predicted_lung_cancer)^2
mse <- mean(squared_errors)
print(mse)

y_2017$error <- with(y_2017, Lung_cancer_count - predicted_lung_cancer)
rmse5 <- mean((y_2017$error/y_2017$Lung_cancer_count)^2)
sqrt(rmse5)
```