environments

MDPI

Article

A Modeling Framework Using Markov Chain and Autologistic Regression to Adjust Temporal and Spatial Dependencies for $PM_{2.5}$ Trajectory Risk Prediction

Rafiqul Chowdhury [1,*] and M. Tariqul Hasan [2,*]

[1] Shannon School of Business, Cape Breton University, Sydney, NS B1M 1A2, Canada
[2] Department of Mathematics and Statistics, University of New Brunswick, Fredericton, NB E3B 5A3, Canada
* Correspondence: rafiqul_chowdhury@cbu.ca (R.C.); thasan@unb.ca (M.T.H.); Tel.: +1-647-204-7234 (R.C.); +1-506-471-3671 (M.T.H.)

Abstract

Data on various air-quality metrics are collected repeatedly by numerous monitoring stations worldwide to closely assess the severity of pollution. Particle pollution from fine particulate matter $PM_{2.5}$ is one such measure used as an indicator of whether air quality is unhealthy. $PM_{2.5}$ is a specific form of air contamination that negatively impacts the environment and human health when levels are above a certain threshold. As data are collected repeatedly over time at multiple locations, there may be a temporal dependence among repeated outcomes and spatial dependence between neighboring stations. Thus, it is important to assess the impact of risk factors on trajectory risk prediction. However, due to the temporal and spatial dependencies, trajectory risk prediction for such data becomes complicated, as both types of dependences must be accounted for during model building. In this paper, we propose a modeling framework that accounts for both types of dependences by incorporating Markov chains and Markov regression, using autologistics for model fitting and trajectory risk prediction. The proposed model fitting and trajectory risk prediction are illustrated using $PM_{2.5}$ outdoor air pollution data from the United States from 2000 to 2020.

Keywords: autologistic; Markov chain; temporal dependence; spatial dependence; longitudinal data; $PM_{2.5}$; trajectory risk

Check for updates

Academic Editors: Benedetto Schiavo, Diana Meza-Figueroa and Claudio Inguaggiato

Received: 15 December 2025
Revised: 27 February 2026
Accepted: 10 March 2026
Published: 12 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Fine particulate matter ($PM_{2.5}$) pollution is typically found in outdoor air due to emissions from the combustion of gasoline, oil, diesel, or wood. Fine particulate matter ($PM_{2.5}$) pollution has been identified as a leading risk factor for the global disease burden [1], as several adverse short- and long-term health impacts have been associated with exposure to $PM_{2.5}$. According to the Environmental Protection Agency (EPA), short-term exposures to $PM_{2.5}$ (up to a 24 h duration) have been associated with premature mortality, increased hospital admissions for heart or lung causes, acute and chronic bronchitis, asthma attacks, emergency room visits, respiratory symptoms, and restricted activity days. These adverse health effects have been reported primarily in infants, children, and older adults with preexisting heart or lung diseases. Long-term (months to years) exposure to $PM_{2.5}$ has been linked to premature death, particularly in people who have chronic heart or lung diseases, and reduced lung function growth in children. In addition, of all of the common air pollutants, $PM_{2.5}$ is associated with the greatest proportion of adverse health effects

related to air pollution, both in the United States and worldwide, based on the World Health Organization's Global Burden of Disease Project. Haze with $PM_{2.5}$ as the primary pollutant has been an atmospheric environmental problem in various parts of the world and has become a significant threat to people's well-being. $PM_{2.5}$ can lead to environmental degradation through reduced visibility, acid rain formation, and ecosystem damage [2]. According to the United States Environmental Protection Agency [3], $PM_{2.5}$ is the main cause of reduced visibility (haze) in parts of the United States, including many treasured national parks and wilderness areas. To be specific, as mentioned by [4], $PM_{2.5}$ poses a serious risk to human health by stimulating oxidative stress and inflammation. Thus, it is very important to perform a rigorous analysis of levels and trends, and risk quantification of $PM_{2.5}$ based on related predictors.

There are varied sources of $PM_{2.5}$. Land use, for example, influences $PM_{2.5}$ levels, with dense vegetation areas tending to have lower concentrations due to the natural air filtration provided by plants [5,6]. Geographically, urban centers often exhibit higher $PM_{2.5}$ levels compared with rural areas due to higher emission sources such as industrial activities and vehicular traffic [7,8]. Additional sources include dust storms, wildfires, and sea spray, as well as anthropogenic activities such as the combustion of fossil fuels and biomass burning [9,10]. Secondary pollutants can also be formed through chemical reactions of primary gaseous emissions [11]. Additionally, meteorological conditions, such as humidity and wind patterns, significantly affect the deposition and dispersion of $PM_{2.5}$ particles, and these conditions are influenced by emission sources [12,13]. This research is motivated by $PM_{2.5}$-based outdoor air pollution data from the United States from 2000 to 2020 and downloaded from the United States Environmental Protection Agency website [3]. The EPA collects data from 1110 sites across the U.S. These data are collected repeatedly at high volume, intensity, and complexity. Our objective is to predict the trajectory risk over the years for given values of available covariates as well as examine how the risk of $PM_{2.5}$ varies across defined risk levels and how controllable levels (categories) of risk factors (sources) affect the trajectories of $PM_{2.5}$. The identification and risk quantification of influencing factors over time are critical for understanding the formation processes and pollution characteristics. However, this task is particularly complex due to temporal and spatial dependencies, with the interplay of multiple factors and their dynamics posing significant challenges. The availability of models that adjust for both temporal and spatial dependences, along with covariates, and accurately predict pollutant patterns across time points is limited [14,15].

To address the modeling of $PM_{2.5}$ data with temporal and spatial dependencies, we propose a covariate-dependent Markov regression model to estimate the impact of risk factors on the trajectory of $PM_{2.5}$ risk. To that end, we propose a joint modeling approach for multiple repeated categorical outcomes at subsequent time points (follow-ups). Note that in this process, one must account for transitions among several states over time, generating a large number of trajectories from the beginning to the end of the study. With the increase in follow-ups, the problem of constructing a joint model for categorical outcomes becomes more difficult and may be unidentifiable. To estimate the impact of risk factors on trajectory risk in repeated-measures data, covariate-dependent Markov regression models are widely used in fields such as rainfall and air pollution modeling. However, most of these applications are used to assess the impact of covariates on transitions [16,17]. The advantage of using the Markov model is that the estimation and model-fitting process is reduced from a multivariate to a univariate approach, due to the effective stratification of the data. The $PM_{2.5}$ data considered in this research exhibit both temporal and spatial dependences, which must be accounted for during model fitting and risk prediction to avoid misleading results. In this research, using repeated

measures and assessing the impact of covariates on transitions, we illustrate trajectory risk prediction to understand air pollution trends by proposing a model that adjusts simultaneously for both temporal and spatial dependences in the data. Efficient prediction of trajectory risk and estimation of the risk of $PM_{2.5}$ exceeding acceptable levels over time will allow policymakers to identify vital information, such as how transitions between states in consecutive responses occur and how covariates influence these transitions. An accurately predicted trajectory will allow us to quantify the risk of air pollution over time, which will be beneficial to identifying factors that may increase air pollution, as required by policymakers.

We start with a brief discussion of the data with the trajectory in Section 2. In Section 3, we proposed a modeling framework based on Markov regression models for the trajectory prediction of repeated categorical responses. The estimation process of the model parameters, along with model significance, is presented in Section 4. An illustrative example using $PM_{2.5}$-based air pollution data is shown in Section 5. We present the conclusions in Section 6.

2. Fine-Particulate-Matter ($PM_{2.5}$)-Based Air Pollution Data

In this paper, we have used fine particulate matter ($PM_{2.5}$) data from 1110 sites across various states in the United States [3]. In $PM_{2.5}$-based air pollution data, the response variable is denoted by Z_{ij} and measured as the yearly maximum value of the $PM_{2.5}$ level, where i refers to the sites and j refers to the years from 2000 to 2020. Following the EPA cutoff for air-quality standards for $PM_{2.5}$ [18], outcomes for all the years are re-categorized as $Z_{ij} = 0$ if $Z_{ij} < 35$ $\mu g/m^3$ and $Z_{ij} = 1$ if $Z_{ij} \geq 35$. Thus, the outcome variable Z_{ij} from the years 2000 to 2020 ($j = 1, \ldots, 21$) for the ith ($i = 1, \ldots, 1110$) site is expressed as $Y_{i1}, \ldots, Y_{i21}$. The frequency distribution of responses ($\mathbf{Z}_{ij}$) for all the years is presented in Table 1.

Table 1. Outcome frequency for the years 2000 to 2020.

Years	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
Z_j	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	Z_7	Z_8	Z_9	Z_{10}	Z_{11}
0	374	363	347	379	354	215	377	217	320	378	301
1	734	708	661	573	492	597	368	372	219	138	181
Total	1108	1071	1008	952	846	812	745	589	539	516	482
Years	2011	2012	2013	2014	12015	2016	2017	2018	2019	2020	
Z_j	Z_{12}	Z_{13}	Z_{14}	Z_{15}	Z_{16}	Z_{17}	Z_{18}	Z_{19}	Z_{20}	Z_{21}	
0	308	319	320	311	292	289	291	256	241	154	
1	103	84	93	109	117	102	93	114	116	40	
Total	411	403	413	420	409	391	384	370	357	194	

It is noteworthy that the monitoring locations (sites) are likely to be spatially correlated, with stronger correlations between sites that are closer. Figure 1 presents the USA map showing the key feature of $PM_{2.5}$ level at the monitoring locations (sites) in various states.

Specifically, we used the proportion of years in which $PM_{2.5}$ exceeded the cutoff points, divided by the total number of years considered in the analysis. In Figure 1, we used various sizes of circles to display the proportion of years that the $PM_{2.5}$ level exceeded cutoff points. For example, the bigger the circle, the higher the proportion of excessive $PM_{2.5}$ levels. We also used various colors to indicate the various proportions. For example, larger circles with the color "yellow" indicate that the proportion is greater than 75% at a specific monitoring site. This means that for 75% of the years considered, the $PM_{2.5}$ level

was higher than the cutoff point at this particular site. Similarly, the colors "blue", "green", and "purple" represent the proportions of 50–74%, 25–49%, and less than 25%, respectively. As mentioned earlier, the $PM_{2.5}$ data are likely to exhibit a spatio-temporal correlation, as the responses are repeatedly collected over time at various locations (monitoring sites). In the next subsection, we verify the presence of spatial correlation.

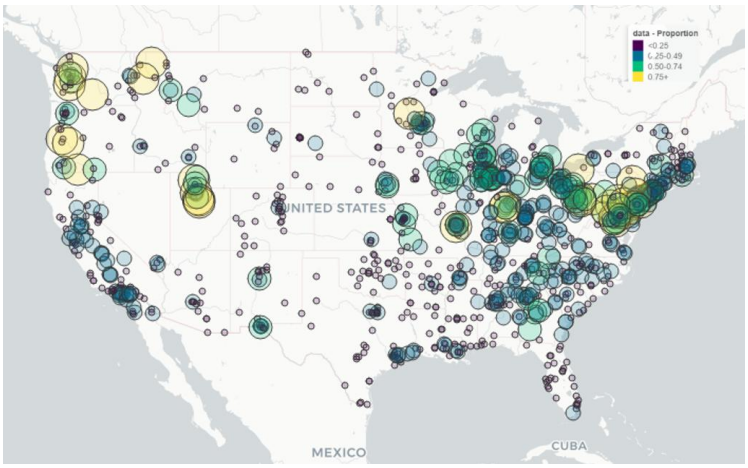


Figure 1. Proportion of $PM_{2.5}$-level-based air-quality indicator at various monitoring sites in the USA.

Moran's I Statistic for Spatial Correlation

To obtain precise estimates and increased accuracy of the risk prediction, it is necessary to investigate the presence of any spatial dependence in the $PM_{2.5}$-related air pollution data. Spatial dependence refers to the extent to which environmental measurements taken at nearby monitoring locations (sites) are statistically correlated. To be specific, whether the $PM_{2.5}$ air pollution measurements at one station are influenced by the same values at surrounding stations, due to the geographical proximity and shared environmental factors.

Analyzing data with spatial dependences often requires spatial statistical methods to account for correlations among nearby locations and to achieve more accurate interpretations of trends and patterns. We need to measure the spatial autocorrelation to describe the extent to which an outcome is correlated with the nearby station's outcome. This phenomenon can be loosely defined as a measure of similarity or dissimilarity between the attribute values of two locations that are spatially proximate. When positive spatial autocorrelation occurs, observations with similar values tend to cluster together. For negative spatial autocorrelation observations, dissimilar values are closer together. Various indices are available to measure the spatial autocorrelation. One most well-known technique is Moran's I statistic [19], used to assess the spatial dependence of whether air-quality ($PM_{2.5}$) measures taken at nearby stations are statistically correlated with each other. For the mathematical formulation of this statistic, readers are referred to [20]. Moran's I statistic is similar to Pearson's correlation, which measures the strength of a linear relationship between two variables [21]. The estimate of Moran's I statistic along with standard errors and p-values for the $PM_{2.5}$ of years 2000 to 2020 are shown in Table 2. Although the autocorrelation values are not strong, ranging from 0.21 to 0.07, the p-values for all years were statistically significant ($p < 0.01$).

Table 2. Estimate with standard errors of the spatial autocorrelation using Moran's I statistic.

Year	Estimates	Standard Errors	*p*-Values
2000	0.115	0.0068	0.000
2001	0.172	0.0063	0.000
2002	0.155	0.0065	0.000
2003	0.189	0.0067	0.000
2004	0.159	0.0071	0.000
2005	0.186	0.0070	0.000
2006	0.167	0.0075	0.000
2007	0.065	0.0231	0.004
2008	0.150	0.0091	0.000
2009	0.153	0.0103	0.000
2010	0.127	0.0109	0.000
2011	0.141	0.0113	0.000
2012	0.096	0.0132	0.000
2013	0.113	0.0134	0.000
2014	0.212	0.0126	0.000
2015	0.135	0.0121	0.000
2016	0.164	0.0124	0.000
2017	0.180	0.0128	0.000
2018	0.133	0.0131	0.000
2019	0.126	0.0136	0.000
2020	0.095	0.0135	0.000

Table 2 indicates the presence of spatial correlation in the $PM_{2.5}$ related air pollution data. Additionally, $PM_{2.5}$ air pollution data are more likely to exhibit a temporal dependence, as responses are collected repeatedly over time. Thus, in the presence of spatio-temporal dependencies in $PM_{2.5}$ air pollution data, we require a modeling technique that can account for both. To that end, in the next section, we propose covariate-dependent Markov regression models to predict trajectory risks that accommodate both spatial and temporal correlations.

3. Trajectory Risk Prediction Using Covariate-Dependent Markov Regression Models

In this section, to analyze $PM_{2.5}$-related air pollution data, we develop covariate-dependent Markov regression models following Chowdhury and Hasan [22]. Let $Y_{i1}, \cdots, Y_{i21}$ represent the binary responses of $PM_{2.5}$-related air pollution levels from the jth ($j = 1, \cdots, 21$) repeated time point for the ith ($i = 1, \cdots, 1110$) site. To illustrate the trajectory, we consider three repeated responses, Y_{i1}, Y_{i2}, and Y_{i3} for the ith site, and $\mathbf{x_{ij}}$ is the vector of covariates for the ith site at the jth year. Henceforth, for simplicity, we omit the covariate subscript. The trajectory is shown in Figure 2. Here, the first column shows marginal probabilities $P(Y_{i1} = y_{i1} \mid \mathbf{x})$, the second column shows first-order conditional probabilities $P(Y_{i2} = y_{i2} \mid y_1; \mathbf{x})$ and the third column shows second-order conditional probabilities $P(Y_{i3} = y_{i3} \mid y_1; y_2; \mathbf{x})$. There are eight possible trajectories. Thus, including many repeated measures (e.g., 2000 to 2020) in the study will require fitting lots of models.

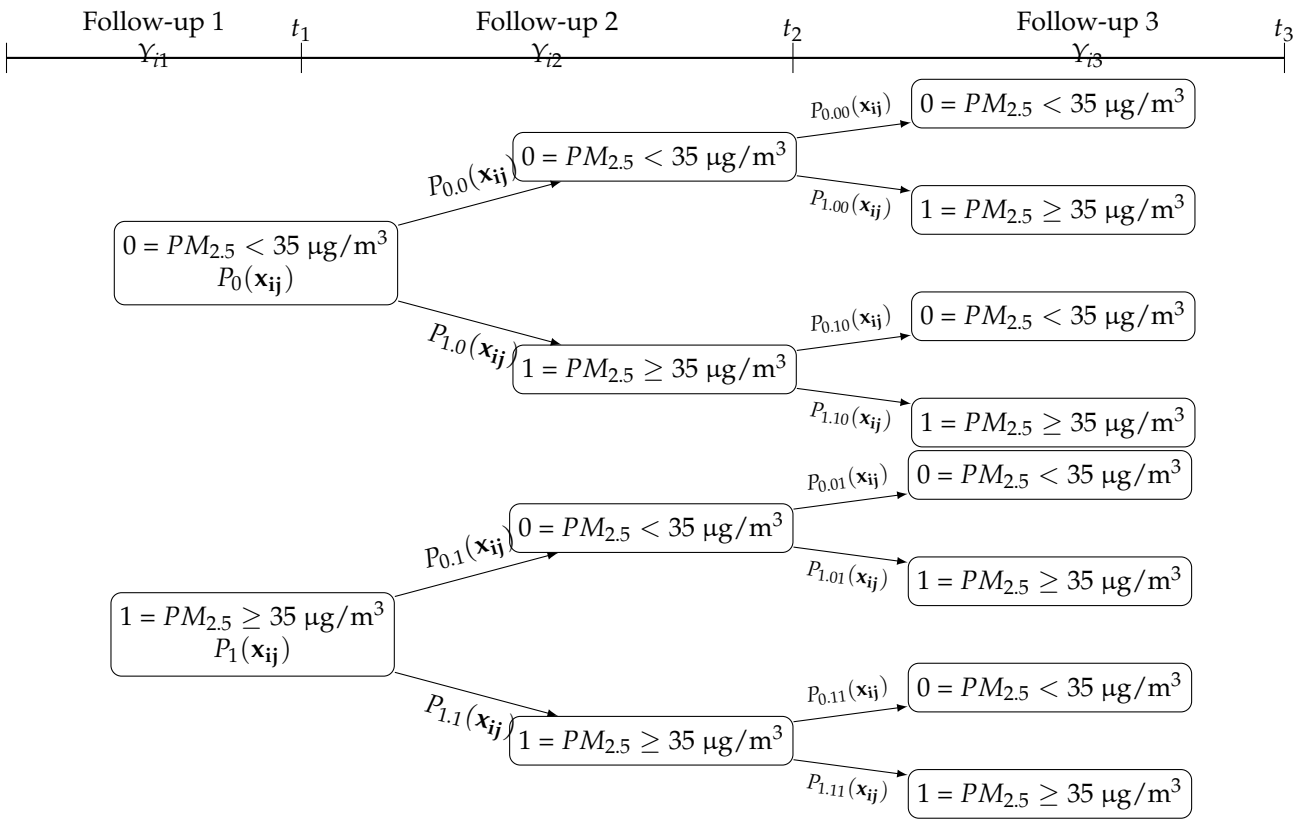


Figure 2. Transitions between two states from three follow-ups.

To obtain trajectory risks, we need to estimate the joint probability, which can be obtained using a Markov chain without relying on a joint model. For example, the outcomes $Y_{i1}, Y_{i2}, \ldots, Y_{ij}$ $(j = 1, \cdots, J)$ in the presence of covariates $\mathbf{x_{ij}}'=(1, x_{i1}, \ldots, x_{ip})$ can be expressed as the product of the marginal and conditional probability mass functions as follows:

$$\begin{aligned} P_{y_{i1},y_{i2},\ldots,y_{ij}}(\mathbf{x_{ij}}) = P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{ij} = y_{ij} \mid \mathbf{x_{ij}}) = \mathbf{P_{y_{i1}}}(\mathbf{x_{ij}}) \\ \times P_{y_{i2}.y_{i1}}(\mathbf{x}_{ij}) \times P_{y_{i3}.y_{i2}y_{i1}}(\mathbf{x}_{ij}) \times \ldots \times P_{y_{ij}.y_{i(j-1)},\ldots,y_{i1}}(\mathbf{x_{ij}}) \end{aligned} \tag{1}$$

which is a Markov chain of order $k(= j - 1)$. In Equation (1), $P_{y_1}(\mathbf{x_{ij}})$ is the marginal probability function of the Y_{i1} conditional on the covariate vector $\mathbf{x_{ij}}$. This is a higher-order Markov model. In other words, the future outcome depends on all previous outcome(s) (higher-order dependence), but this would require a very large sample size for each time point.

The trajectory risk using Equation (1) with three time points shown in Figure 1 would be as follows.

$$\begin{aligned} P_{y_{i1},y_{i2},y_{i3}}(\mathbf{x_{ij}}) &= P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, Y_{i3} = y_{i3} \mid \mathbf{x_{ij}}) \\ &= P_{y_{i1}}(\mathbf{x}_{ij}) \times P_{y_{i2}.y_{i1}}(\mathbf{x}_{ij}) \times P_{y_{i3}.y_{i2}y_{i1}}(\mathbf{x_{ij}}) \end{aligned} \tag{2}$$

where $P(Y_{i2} = y_{i2} \mid y_1; \mathbf{x}) = P_{y_{i2}.y_{i1}}(\mathbf{x}_{ij}), P(Y_{i3} = y_{i3} \mid y_1; y_2; \mathbf{x}) = P_{y_{i3}.y_{i2}y_{i1}}(\mathbf{x_{ij}})$ and so on. This is a second-order Markov chain as outcome Y_{i3} depends on the previous two outcomes, Y_{i2} and Y_{i1}, respectively.

3.1. Covariate-Dependent Markov Regression Models

The Markov model is a common choice for studying covariate impacts on transition probabilities used for risk prediction, considering the underlying dependence in consecutive

time points [16,17,23]. A natural choice is logistic regression to model the transition ($Y_{ij} \rightarrow Y_{i(j+1)}$) as a function of the covariate vector $\mathbf{x}_{ij}$ for the ith site. Then, the first order conditional probability Y_{i2} given Y_{i1} and $\mathbf{x}_{ij}$ (the probability mass function of the Y_{i2} conditional on Y_{i1} and $\mathbf{x_{ij}}$) can be expressed as $P(Y_{i2} = y_{i2} \mid Y_{i1} = y_{i1}, \mathbf{x}_{ij})$, stratifying the sample based on $Y_{i1} = 0$ and $Y_{i1} = 1$ as the outcome Y_{i1} is binary. To be specific, when we stratify the data using $Y_{i1} = 0$ and model Y_{i2} as a function of the covariates, we are modeling the transitions from $Y_{i1} = 0$ to $Y_{i2} = s$ (i.e., $Y_{i1} = 0 \rightarrow Y_{i2} = s$) where $s = 0, 1$. Similarly, when we stratify the data using $Y_{i1} = 1$ and model Y_{i2} as a function of covariates, we are modeling the transitions ($Y_{i1} = 1 \rightarrow Y_{i2} = s$). Then, the impact of the covariates on the first order Markov model for $P(Y_{i2} = y_{i2} \mid Y_{i1} = 0, \mathbf{x}_{ij})$ can be shown as:

$$P_{y_{i2}.0}(\boldsymbol{x}_{ij}) = P(Y_{i2} = s \mid Y_{i1} = 0, \boldsymbol{x}_{ij}) = \frac{e^{(\beta_{00}+\beta_{01}x_{i1}+\ldots+\beta_{0p}x_{ip})s}}{1+e^{(\beta_{00}+\beta_{01}x_{i1}+\ldots+\beta_{0p}x_{ip})}} = \frac{e^{(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{y_2.0})s}}{1+e^{(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{y_2.0})}}, s = 0, 1, \tag{3}$$

where $\boldsymbol{\beta}'_{y_2.0} = [\beta_{00}, \beta_{01}, \ldots, \beta_{0p}]$ is a $1 \times (p+1)$ vector of regression coefficients corresponding to the covariate vector $\mathbf{x'_{ij}} = [\mathbf{1}, \mathbf{x_{i1}}, \ldots, \mathbf{x_{ip}}]$ for the sub-sample stratified on $Y_{i1} = 0$.

Similarly, the first order conditional model for $P(Y_{i2} = y_{i2} \mid Y_{i1} = 1, \mathbf{x_{ij}})$ is:

$$P_{y_{i2}.1}(\boldsymbol{x}_{ij}) = P(Y_{i2} = s \mid Y_{i1} = 1, \boldsymbol{x}_{ij}) = \frac{e^{(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{\mathbf{y_2.1}})s}}{1+e^{(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{\mathbf{y_2.1}})}}, \quad s = 0, 1, \tag{4}$$

where $\boldsymbol{\beta}_{\mathbf{y_2.1}}{}' = [\beta_{10}, \beta_{11}, \ldots, \beta_{1p}]$ is a $1 \times (p+1)$ vector of regression coefficients corresponding to the covariate vector $\mathbf{x'_{ij}}$ for the sub-sample stratified on $Y_{i1} = 1$.

Consequently, Equations (3) and (4), $P(Y_{i2} = s \mid Y_{i1} = y_{i1}, \boldsymbol{x}_{ij})$ can be simplified as :

$$P_{y_{i2}.y_1}(\boldsymbol{x}_{ij}) = P(Y_{i2} = s \mid Y_{i1} = y_{i1}, \boldsymbol{x}_{ij}) = \frac{e^{(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{\mathbf{y_2.y_1}})s}}{1+e^{(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{\mathbf{y_2.y_1}})}}, \quad s, y_1 = 0, 1, \tag{5}$$

where $\boldsymbol{\beta}_{\mathbf{y_2.y_1}}{}' = [\beta_{y_10}, \beta_{y_11}, \ldots, \beta_{y_1p}]$ is a $1 \times [2(p+1)]$ vector of regression coefficients corresponding to the covariate vector $\boldsymbol{x}'_{ij} = [1, x_1, \ldots, x_p]$ for the sub-sample stratified on $Y_{i1} = y_{i1}$. This first-order Markov model easily generalizes to a higher-order [22].

The $(j-1)$ *th* order Markov models shown in (1) is conditional on all previous histories. These models require large samples at every time point for model fitting. If the sample size is inadequate, one can assess only first-order dependencies and predict trajectory risks—for example, the air-quality data. In this case, we can develop models by considering pairwise (first-order) conditional models. The first-order Markov models for j consecutive outcomes $Y_{i1}, \cdots, Y_{ij}$ can be shown as follows:

$$P_{y_{ij}.y_{j-1}}(\boldsymbol{x}_{ij}) = P(Y_{ij} = s \mid Y_{i(j-1)} = y_{i(j-1)}, \boldsymbol{x}_{ij}) = \frac{e^{(\boldsymbol{x}'_{ij}\boldsymbol{\beta}_{\mathbf{y_j.y_{j-1}}})s}}{1+e^{(\mathbf{x'_{ij}}\boldsymbol{\beta}_{\mathbf{y_j.y_{j-1}}})}}, \quad s, y_{i(j-1)}, y_{ij} = 0, 1, \tag{6}$$

where $\boldsymbol{\beta}_{\mathbf{y_j.y_{j-1}}}{}' = [\beta_{y_10}, \beta_{y_11}, \ldots, \beta_{y_1p}]$ is a $1 \times [2(p+1)]$ vector of regression coefficients corresponding to the covariate vector $\boldsymbol{x}'_{ij} = [1, x_1, \ldots, x_p]$ for the sub-sample stratified on $Y_{i(j-1)} = y_{i(j-1)}$.

The trajectory risk based on consecutive first-order Markov models (6) can be obtained using the following Markov chain:

$$P_{y_{i1}, y_{i2}, \ldots, y_{iJ}}(\boldsymbol{x_{ij}}) = P_{is}(\boldsymbol{x}_{ij}) \times P_{y_{i2}.y_{i1}}(\boldsymbol{x}_{ij}) \times P_{y_{i3}.y_{i2}}(\boldsymbol{x_{ij}}) \times \ldots \times P_{y_{ij}.y_{i(j-1)}}(\boldsymbol{x}_{ij}). \tag{7}$$

We can obtain the first-order conditional probabilities for all consecutive outcomes using Equation (6).

3.2. Markov Chain and Temporal Dependency

The Markov chain has a built-in mechanism to handle temporal dependency. To predict the trajectory risk, we need to estimate the joint probability. In Markov chain, the joint probability (trajectory risk) for a site with a specified covariate vector $(\mathbf{x_{ij}}^*)$ for a trajectory $\hat{P}(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{ij} = y_{ij} \mid \mathbf{x_{ij}}^*)$ is predicted using the following $k = (j\text{-}1)$th order Markov chain:

$$\begin{aligned}\hat{P}(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \ldots, Y_{iJ} = y_{iJ} \mid \boldsymbol{x_{ij}}) &= \hat{P}(Y_{i1} = y_{i1} \mid \boldsymbol{x_{ij}}) \times \hat{P}(Y_{i2} = y_{i2} \mid Y_{i1} = y_{i1}; \boldsymbol{x_{ij}}) \\ &\times \hat{P}(Y_{i3} = y_{i3} \mid Y_{i2} = y_{i2}, Y_{i1} = y_{i1}; \boldsymbol{x_{ij}}) \times, \ldots, \times \hat{P}(Y_{iJ} = y_{iJ} \mid Y_{i(j-1)} = y_{i(j-1)}, \ldots, Y_{i1} = y_{i1}; \boldsymbol{x_{ij}}) \\ &= \hat{P}_{y_{i1}}(\boldsymbol{x_{ij}}) \times \hat{P}_{y_{i2}.y_{i1}}(\boldsymbol{x}_{ij}) \times \hat{P}_{y_{i3}.y_{i2}.y_{i1}}(\boldsymbol{x}_{ij}) \times \ldots \times \hat{P}_{y_{iJ}.y_{i(j-1)},\ldots,y_{i1}}(\mathbf{x_{ij}}).\end{aligned} \tag{8}$$

The left-hand side of Equation (8), which is the joint probabilities and what we want to estimate, is denoted as $\hat{P}_{y_{i1},y_{i2},\ldots,y_{iJ}}(\mathbf{x_{ij}})$. If available, we could predict the above joint probabilities in (8) for trajectory risk using a joint model for repeated binary outcomes that considers the effect of covariates. However, no such models for many repeated binary outcomes (e.g., multivariate logistic regression) are available.

The first term of the right-hand side of Equation (8), $\hat{P}(Y_{i1} = y_{i1} \mid \boldsymbol{x_{ij}}) = \hat{P}_{y_{i1}}(\mathbf{x_{ij}})$ is the marginal probabilities for different values of covariate vectors $\mathbf{x_{ij}}$ and $\mathbf{y_{ij}} = \mathbf{0}, \mathbf{1}$. These probabilities can be estimated using the fitted marginal model using Equation (5) and the data from the first time point. In other words, these provide the initial probabilities for the Markov chain.

The second term of the right-hand side of Equation (8), $\hat{P}(Y_{i2} = y_{i2} \mid Y_{i1} = y_{i1}; \boldsymbol{x_{ij}}) = \hat{P}_{y_{i2}.y_{i1}}(\boldsymbol{x}_{ij})$ are the conditional probabilities for different values of the covariates vector $\mathbf{x_{ij}}$ and for $\mathbf{y_{ij}} = \mathbf{0}, \mathbf{1}$ and so on. Again, these probabilities can be estimated using the fitted higher-order Markov models. It should be noted that to estimate first-order probabilities, we use two models for $y_{i1} = 0$ and $y_{i1} = 1$ as shown in Equations (1) and (2), respectively. Since the first-order Markov model uses two consecutive binary outcomes y_{ij} and $y_{i(j-1)}$, we need to fit two separate models.

Then, for outcomes y_{i1}, y_{i2} with categories 0,1 and using Equation (8), we can predict joint probabilities for all possible values of y_{i1}, y_{i2} and x_{ij} using marginal and conditional probabilities as follows:

$$\begin{aligned}\hat{P}(Y_{i1} = y_{i1}, Y_{i2} = y_{i2} \mid \boldsymbol{x_{ij}}) = \hat{P}(Y_{i1} = y_{i1}; \boldsymbol{x_{ij}}) \times \hat{P}(Y_{i2} = y_{i2} \mid Y_{i1} = y_{i1}; \boldsymbol{x_{ij}}), \\ y_{i1} = 0, 1;\ y_{i2} = 0, 1;\ j = 1, 2.\end{aligned}$$

Similarly, for all j outcomes $y_{i1}, y_{i2}, \ldots, y_{ij}$, we can predict joint probabilities. For example, $\hat{P}_{0,1}(\boldsymbol{x}^*) = \hat{P}_0(\boldsymbol{x}^*) \times \hat{P}_{1.0}(\boldsymbol{x}^*)$ and $\hat{P}_{0,0}(\boldsymbol{x}^*) = 1 - \hat{P}_{0,1}(\boldsymbol{x}^*)$ and so on. It may be noted that for simplicity, the subscript i is omitted from the predicted joint probabilities $P_{0,1}(\mathbf{x}^*)$ and $P_{0,0}(\mathbf{x}^*)$. Therefore, we will obtain these two probabilities for each site.

It is clear from Equations (1) and (2) that these two models are univariate because the yearly repeated data were stratified cross-sectionally by the previous years' outcome categories. In other words, all datasets from 2 onward are cross-sectional (univariate) and are used to estimate Markov regression models for conditional probabilities by accounting for temporal dependence. We use Equation (8) to obtain the joint probabilities from the univariate probabilities. Hence, the Markov chain addresses the problem of temporal dependence.

3.3. Adjustment of Spatial Dependence Using Autologistic Regression

In this section, we extend the proposed covariate-dependent Markov regression model to adjust for spatial dependency. It should be noted that the logistic regression for the binary outcome we adopted earlier does not account for spatial dependence. As our objective is to predict trajectory risk and obtain more precise and accurate estimates, we are interested in improvements from adjusting spatial and temporal dependencies.

Logistic regression is the most widely used model for analyzing binary responses in the presence of covariates; however, it does not directly account for spatial dependences in the data. The autologistic regression model incorporates a spatially lagged term to account for the behavior of neighboring agents, thereby adjusting for spatial dependences in the data [24]. Autologistic examples can be seen in ecology [25–28], health [29,30], atmospheric science [31], and mining [32,33] among others.

In the autologistic regression model, an additional explanatory variable is added, along with the existing covariates, to capture the effect of other response variables in the spatial neighborhood. Using the form of the ordinary logistic regression model shown in Equation (6), the autologistic regression model can be shown by adding a spatially lagged term accounting for spatial autocorrelation as follows:

$$\begin{aligned} P_{y_{ij}.y_{j-1}}(\boldsymbol{z}_{ij}) &= P_{y_{ij}.y_{j-1}}(\boldsymbol{x}_{ij}; A_{ij}) = P(Y_{ij} = s \mid Y_{i(j-1)} = y_{i(j-1)}, \boldsymbol{x}_{ij}, A_{ij}) \\ &= \frac{e^{(\beta_{00}+\beta_{01}x_{i1}+\ldots+\beta_{0p}x_{ip}+\gamma_{0(p+1)}A_{ij})s}}{1+e^{(\beta_{00}+\beta_{01}x_{i1}+\ldots+\beta_{0p}x_{ip}+\gamma_{0(p+1)}A_{ij})}} = \frac{e^{(\boldsymbol{z}'_{ij}\boldsymbol{\alpha}_{\boldsymbol{y_j.y_{j-1}}})s}}{1+e^{(\boldsymbol{z}'_{ij}\boldsymbol{\alpha}_{\boldsymbol{y_j.y_{j-1}}})}}, \quad s, y_{i(j-1)}, y_{ij} = 0, 1, \end{aligned} \tag{9}$$

where $\boldsymbol{\alpha}_{\boldsymbol{y_j.y_{j-1}}}{}' = \left[\beta_{y_1 0}, \beta_{y_1 1}, \ldots, \beta_{y_1 p}, \gamma_{0(p+1)}\right]$ is a $1 \times [2(p+1)]$ vector of regression coefficients corresponding to the covariate vector $\boldsymbol{x}'_{ij} = \left[1, x_1, \ldots, x_p, A_{p+1}\right]$ for the sub-sample stratified on $Y_{i(j-1)} = y_{i(j-1)}$; A_i is the spatially lagged explanatory variable (also known as autocovariate) and $\gamma_{0(p+1)}$ is the unknown parameter corresponding to A_i that needs to be estimated. This autocovariate term (A_i), used by others [24,34–37], can be estimated as follows:

$$A_i = \frac{\sum_{\substack{i=1 \\ i \neq i'}}^{k} W_{ii'} \hat{p}_{i'}}{\sum_{\substack{i=1 \\ i \neq i'}}^{k} W_{ii'}}.$$

where $W_{ii'}$ is the spatial weight between monitoring stations i and i', and $\hat{p}$ is the predicted probability at monitoring station i by the logistic regression model. k represents the number of neighbouring monitoring stations around i to be included in the calculation of the A_i term. The trajectory risk based on consecutive first-order Markov models (9) can be obtained using Equation (7) as before.

4. Parameter Estimation

The log-likelihood function of the joint mass function for the first-order Markov chain in (7) can be shown to be:

$$\begin{aligned} l &= \sum_{i=1}^{n}\sum_{j=1}^{J} \ln P(Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \cdots, Y_{iJ} = y_{iJ} \mid \boldsymbol{x_{ij}}) = \sum_{i=1}^{n}\sum_{j=1}^{J}\Big[\ln P(Y_{i1} = y_{i1} \mid \boldsymbol{x_{ij}}) \\ &+ \ln P(Y_{i2} = y_{i2} \mid y_{i1}; \boldsymbol{x_{ij}}) + \cdots + \ln P(Y_{iJ} = y_{iJ} \mid y_{i(j-1)}; \boldsymbol{x_{ij}})\Big] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J}\Big[\ln P_{y_1}(\boldsymbol{x_{ij}}) + \ln P_{y_2.y_1}(\boldsymbol{x_{ij}}) + \cdots + \ln P_{y_{ij}.y_{j-1}}(\boldsymbol{x_{ij}})\Big]. \end{aligned} \tag{10}$$

From Equation (10), it is clear that the joint mass function reduces to the univariate mass function for a single outcome variable. In other words, all the consecutive first-order

conditional models are equivalent to that of the marginal model (univariate logistic or auto logistic regression). All these models can be estimated in a standard manner using an appropriate data structure and Python 3.14.3, R 4.3.3, or other software capable of fitting these models. Then, we can exploit parallel programming to analyze all data subsets using multiple cores on a single computer or several CPUs in a distributed system.

The complete R code to fit models simultaneously on 41 datasets using autologistic regression and logistic regression, including predicted probability, is shown in Appendix A, Table A9. This code also included all the necessary R packages used for our analysis. We used a parallel programming approach using multiple processors to fit all the models.

5. Analysis of Fine Particulate Matter ($PM_{2.5}$) Data

5.1. Data Preparation

To fit the Markov regression model, we must stratify the data by transitions between consecutive years. For the years 2000 to 2020, there will be 41 datasets (requiring the fitting of 41 separate models), numbered from 1 to 41. The dataset for the year 2000 (marginal model) is assigned codes 1 and 2 to 41 for the two types of transitions. For example, code 2 represents the dataset for transitions $Y_{i1} = 0 \rightarrow Y_{i2} = s$, code 4 represents the dataset for transitions $Y_{i2} = 0 \rightarrow Y_{i3} = s$ as shown in Table 3. We assigned even numbers for datasets for any transition of type $0 \rightarrow s$ where $(s = 0, 1)$, whereas the number 3 represents the dataset for transitions $Y_{i1} = 1 \rightarrow Y_{i2} = s$ $(s = 0, 1)$. The number 5 represents the dataset for transitions $Y_{i2} = 1 \rightarrow Y_{i3} = s$ $(s = 0, 1)$. That is, we assigned odd numbers for datasets for any transition of type $1 \rightarrow s$ where $(s = 0, 1)$. By creating transition indicators, we stratify all observations for the current year by the two categories (0, 1) of the previous year. Table 3 presents the datasets and transition (dependent) variables constructed from several hypothetical observations.

Datasets 2–41 (Table 4) can be used to fit first-order Markov regression models (conditional models) to estimate conditional probabilities. Because of the small sample size (1110 sites), it may not be possible to include higher-order terms, as there may not be sufficient cell frequencies to fit the model. Table 4 also presents the frequency distribution for all 41 datasets. The description of the covariates used in the model is shown in Table 5. The variables included in the augmented data file, and the data file structure are shown in Table 6. The Moran's I statistics for all datasets from the consecutive transitions are shown in Table 7. In general, Moran's I values are lower across all datasets.

Table 3. Transitions indicator with dependent variable based on the four hypothetical sites.

Site (i)	Year (j)	Year (j+1)	Z_{ij}	$Z_{i(j+1)}$	Transitions	Depvar (Y)	Datasets
1	2000	2001	0	0	$0 \rightarrow 0$	0	2
2	2000	2001	0	1	$0 \rightarrow 1$	1	2
3	2000	2001	1	1	$1 \rightarrow 1$	0	3
4	2000	2001	1	0	$1 \rightarrow 0$	1	3
1	2001	2002	0	0	$0 \rightarrow 0$	0	4
2	2001	2002	0	1	$0 \rightarrow 1$	1	4
3	2001	2002	1	1	$1 \rightarrow 1$	0	5
4	2001	2002	1	0	$1 \rightarrow 0$	1	5
1	2002	2003	0	0	$0 \rightarrow 0$	0	6
2	2002	2003	1	0	$1 \rightarrow 0$	1	7
3	2002	2003	1	1	$1 \rightarrow 1$	0	7
4	2002	2003	1	0	$1 \rightarrow 0$	1	7

Table 4. Outcome frequency for the subsets for two transition types and the marginal model.

Subset and Transition Type: $0 \rightarrow 1$													
Model (Dataset) Number													
	2	4	6	8	10	12	14	16	18	20	22	24	26
Y_j	Y_2	Y_4	Y_6	Y_8	Y_{10}	Y_{12}	Y_{14}	Y_{16}	Y_{18}	Y_{20}	Y_{22}	Y_{24}	Y_{26}
0	235	213	213	209	159	145	161	164	265	265	200	253	279
1	121	109	98	93	173	33	130	31	38	88	43	27	31
Total	356	322	311	302	332	178	291	195	303	353	243	280	310
Continued... Subset and Transition type: $0 \rightarrow 1$													
Model (Dataset) Number													
	28	30	32	34	36	38	40						
Y_j	Y_{28}	Y_{30}	Y_{32}	Y_{34}	Y_{36}	Y_{38}	Y_{40}						
0	266	254	231	244	236	208	128						
1	41	40	46	36	42	36	13						
Total	307	294	277	280	278	244	141						
Subset and Transition type: $1 \rightarrow 1$													
Model (Dataset) Number													
	3	5	7	9	11	13	15	17	19	21	23	25	27
Y_j	Y_3	Y_5	Y_7	Y_8	Y_{11}	Y_{13}	Y_{15}	Y_{17}	Y_{19}	Y_{21}	Y_{23}	Y_{25}	Y_{27}
0	130	135	167	147	55	232	55	152	109	34	104	50	25
1	587	552	475	399	422	335	238	187	99	93	58	49	58
Total	717	687	642	546	477	567	293	339	208	127	162	99	83
Continued... Subset and Transition type: $1 \rightarrow 1$										Marginal model			
Model (Dataset) Number													
	29	31	33	35	37	39	41			1			
Y_j	Y_{29}	Y_{31}	Y_{33}	Y_{35}	Y_{37}	Y_{39}	Y_{41}			Y_1			
0	23	31	57	47	19	32	26			376			
1	67	76	55	54	70	80	27			734			
Total	90	107	112	101	89	112	53			1110			

Table 5. Descriptions of covariates used in the models.

Variables	**Description**	**Values**
SiteID	ID number for each monitor site	
Event (X_1)	Occurance of a wildfire	Yes = 1, No = 0
Elevation (X_2)	Measured in meters	Ranges between 0 to 7085
Area (X_3)	Location of the monitors	Urban/Suburban = 1, Rural/Others = 0
Land (X_4)	Type of land	Commercial/Industrial/ Agriculture = 1, Residential/Others = 0
DivID	Dataset identification number	1 to 41
Latitude (X_5)	Site Latirude	
Longitude (X_6)	Site Longitude	

Missing Data Handling

Missing data are common in longitudinal and time-series data for various reasons. The same is true for environmental data; however, it is less prone to error because it is captured automatically via monitors. However, in our proposed modeling framework, data stratification reduces a multivariate problem to a univariate case. Therefore, one can easily apply all appropriate techniques for handling missing data as we would do for a single dataset. Additionally, the presence of outliers in this dataset does not pose a threat to model building, as shown in the covariate details in Table 5. It is strictly categorical except for one. Also, the proposed framework can accommodate a varying number of covariates per stratified dataset, which can be handled through additional programming tasks.

5.2. Analysis Results

The proposed methodology is illustrated by applying it to fine-particulate-matter (PM2.5)-based air pollution data collected from 1110 sites across various states in the United States. We fitted three sets of models to all datasets. The following are the models:

(i) Logistic Model 1: This model includes six independent variables from X_1 to X_6 as identified in Table 5.

(ii) Logistic Model 2: This model includes four independent variables from X_1 to X_4 as identified in Table 5.

(iii) Autologistic: This model includes four independent variables from X_1 to X_4 and an additional variable, a spatially lagged term that accounts for the spatial autocorrelation derived based on latitude (X_5) and longitude (X_6) variables as identified in Table 5.

(iv) Logistic Model 4: In this model, we are assuming no temporal dependence. This implies that we can fit a total of 21 logistic regression models, one for each year. We do not need the data stratification that will create a total of 41 datasets. However, if the temporal dependence is there, our standard error of regression coefficients will be distorted, hence distorting inferences. In addition, it may affect the trajectory risk prediction.

If the logistic regression model 4 holds, then Equation (8) becomes as follows:

$$\begin{aligned}&\hat{P}(Y_{i1}=y_{i1},Y_{i2}=y_{i2},\ldots,Y_{iJ}=y_{iJ}\mid \boldsymbol{x_{ij}})=\hat{P}(Y_{i1}=y_{i1}\mid \boldsymbol{x_{ij}})\times\hat{P}(Y_{i2}=y_{i2}\mid \mathbf{x_{ij}})\\&\times,\ldots,\times\hat{P}(Y_{iJ}=y_{iJ}\mid \boldsymbol{x_{ij}})=\hat{P}_{y_{i1}}(\boldsymbol{x_{ij}})\times\hat{P}_{y_{i2}}(\boldsymbol{x}_{ij})\times\ldots\times\hat{P}_{y_{ij}}(\boldsymbol{x}_{ij}),\ \text{where } y_{iJ}=0,1.\end{aligned}\tag{11}$$

Table 6. Augmented data file structure.

SiteID	Y	Event	Elevation	Area	Land	DivID	Latitude	Longitude
10030010	1	0	37.19	1	1	1	30.50	−87.88
10030010	0	0	37.19	1	1	4	30.50	−87.88
10030010	0	1	37.19	1	1	8	30.50	−87.88
10030010	0	0	37.19	1	1	12	30.50	−87.88
10030010	0	0	37.19	1	1	3	30.50	−87.88
10030010	0	0	37.19	1	1	6	30.50	−87.88
10030010	0	0	37.19	1	1	10	30.50	−87.88
10270001	1	0	344.42	0	0	4	33.28	−85.80
10270001	1	0	344.42	0	0	10	33.28	−85.80
10270001	1	0	344.42	0	0	1	33.28	−85.80
10270001	0	0	344.42	0	0	13	33.28	−85.80

Table 6. *Cont.*

SiteID	Y	Event	Elevation	Area	Land	DivID	Latitude	Longitude
10270001	0	0	344.42	0	0	9	33.28	−85.80
10270001	1	0	344.42	0	0	7	33.28	−85.80
10270001	0	0	344.42	0	0	3	33.28	−85.80
10331002	1	0	122.00	1	1	4	34.76	−87.64
10331002	0	0	122.00	1	1	13	34.76	−87.64
10331002	1	0	122.00	1	1	1	34.76	−87.64
10331002	0	0	122.00	1	1	3	34.76	−87.64
10331002	1	0	122.00	1	1	7	34.76	−87.64
10331002	0	0	122.00	1	1	9	34.76	−87.64

Table 7. Moran's I statistic of spatial autocorrelation for all new datasets.

Datasets	Observed	Expected	sd	p.v	Datasets	Observed	Expected	sd	p.v
1	0.115	−0.0009	0.0068	0.000	21	0.080	−0.0079	0.0342	0.010
2	0.157	−0.0028	0.0141	0.000	22	0.071	−0.0041	0.0177	0.000
3	0.118	−0.0014	0.0088	0.000	23	0.164	−0.0062	0.0315	0.000
4	0.128	−0.0031	0.0153	0.000	24	0.094	−0.0036	0.0157	0.000
5	0.093	−0.0015	0.0091	0.000	25	0.181	−0.0102	0.0424	0.000
6	0.140	−0.0032	0.0159	0.000	26	0.204	−0.0032	0.0143	0.000
7	0.170	−0.0016	0.0095	0.000	27	0.339	−0.0122	0.0536	0.000
8	0.164	−0.0033	0.0173	0.000	28	0.063	−0.0033	0.0139	0.000
9	0.111	−0.0018	0.0097	0.000	29	0.104	−0.0112	0.0535	0.032
10	0.246	−0.0030	0.0141	0.000	30	0.138	−0.0034	0.0127	0.000
11	0.104	−0.0021	0.0117	0.000	31	0.145	−0.0094	0.0443	0.000
12	0.150	−0.0057	0.0286	0.000	32	0.223	−0.0036	0.0149	0.000
13	0.170	−0.0018	0.0089	0.000	33	0.335	−0.0090	0.0392	0.000
14	0.178	−0.0034	0.0165	0.000	34	0.056	−0.0036	0.0159	0.000
15	0.080	−0.0034	0.0162	0.000	35	0.330	−0.0100	0.0412	0.000
16	0.126	−0.0052	0.0210	0.000	36	0.073	−0.0036	0.0153	0.000
17	0.121	−0.0030	0.0150	0.000	37	0.143	−0.0114	0.0410	0.000
18	0.044	−0.0033	0.0164	0.004	38	0.057	−0.0041	0.0177	0.001
19	0.149	−0.0048	0.0235	0.000	39	0.177	−0.0090	0.0354	0.000
20	0.097	−0.0028	0.0142	0.000	40	0.030	−0.0071	0.0304	0.220

Then, the following form of autologistic regression can be used to assess the impact of covariates:

$$P_{y_{ij}}(\boldsymbol{z}_{ij}) = P_{y_{ij}}(\boldsymbol{x}_{ij}; A_{ij}) = P(Y_{ij} = s \mid \boldsymbol{x}_{ij}, A_{ij}) = \frac{e^{(\beta_{00}+\beta_{01}x_{i1}+\ldots+\beta_{0p}x_{ip}+\gamma_{0(p+1)}A_{ij})s}}{1+e^{(\beta_{00}+\beta_{01}x_{i1}+\ldots+\beta_{0p}x_{ip}+\gamma_{0(p+1)}A_{ij})}} = \frac{e^{(\boldsymbol{z}'_{ij}\boldsymbol{\alpha}_{y_j.y_{j-1}})s}}{1+e^{(\boldsymbol{z}'_{ij}\boldsymbol{\alpha}_{y_j.y_{j-1}})}}, \quad s, y_{i(j-1)}, y_{ij} = 0, 1, \tag{12}$$

where $\boldsymbol{\alpha}_{y_j.y_{j-1}}{}' = \left[\beta_{y_1 0}, \beta_{y_1 1}, \ldots, \beta_{y_1 p}, \gamma_{0(p+1)}\right]$ is a $1 \times [(p+1)]$ vector of regression coefficients corresponding to the covariate vector $\boldsymbol{x}'_{ij} = \left[1, x_1, \ldots, x_p, A_{p+1}\right]$ for the sample data from each year; A_i is the spatially lagged explanatory variable (also known as autocovariate) and $\gamma_{0(p+1)}$ is the unknown parameter corresponding to A_i that needs to be estimated. The trajectory risk can be obtained using Equation (11) as before.

The proposed framework is based on a statistical approach to analyzing these longitudinal datasets. The covariate impact on binary outcomes is estimated using logistic regression and autologistic regression. The linking of repeated-measures data was performed using joint, marginal, and conditional models by data stratification. The longitudinal data allows us to observe how individual responses change over time, which is very useful for trajectory analysis [38]. This is generally done by conditioning on a subject's previous

history. Lee and Nelder [39] concluded that conditional models were of fundamental interest and that marginal predictions could be derived from them. Of course, different machine-learning (ML) models can be used for risk prediction. However, it is beyond the scope and objectives of this paper, as we are interested in hypothesis testing, statistical inference, and trajectory risk prediction. ML models do not align with all of our objectives. Chowdhury and Islam [40] proposed predicting joint probabilities by linking marginal and conditional probabilities in a Markov chain for multinomial outcomes.

Table 8 presents the likelihood ratio (L.R.) statistics and Akaike information criterion (AIC) for the above three models and all the datasets. The L.R. and AIC statistics for the autologistic models across all datasets are the lowest, indicating that the autologistic models are the best-fitted. The prediction accuracies and the area under the curve (AUC) for all three models across all datasets are shown in Table 9. In this table, the last two columns present the accuracy for independent logistic (Logistic model 4) and autologistic regression models for individual yearly datasets, that is, 2000–2019 (20 sets) . This implies that we assume temporal independence across the years. The independence models achieve the lowest accuracy. However, compared with Model 1, the autologistic regression models across all datasets achieve the highest accuracies and are therefore the best-fitting models, as shown in Table 8. In general, the autologistic models perform better, as measured by AUC, than other models. Also, the relative operative characteristic (ROC) curves provide very similar findings, as shown in the Appendix A, Figures A1–A3. In other words, we need to use models that adjust for both temporal and spatial dependences in the data. For example, for trajectory risk prediction and statistical inference, we need to use the autologistic model proposed within the Markov regression framework.

The directions of the relationships between the covariates and the dependent variables, along with the significance levels for the autologistic regression models, are shown in Table 10. The coefficients on the spatial dependence variable (AutoCov) across all Markov models are positive and highly significant ($p < 0.01$). However, in general, the coefficient sizes for models that transitioned from $1 \rightarrow 1$ (models with odd numbers) were larger than those for models that transitioned from $0 \rightarrow 1$ (models with even numbers), as expected. Other covariates (X_1 to X_4) showed both positive and inverse relationships with the dependent variables, with some significant at the 5% or 1% levels and others nonsignificant. For comparison, the directions of the relationships between the covariates and the dependent variable, along with their significance levels, for all independent logistic regression models, are shown in Table 11. While the significance levels and directions for the AutoCov variables remain the same, the direction and significance level of the relationship showed some changes.

Table 8. Model statistics for different logistic and autologistic models for all datasets.

Data	Logistic						Data	Logistic					
	Model 1		Model 2		Auto			Model 1		Model 2		Auto	
Sets	L.R.	AIC	L.R.	AIC	L.R.	AIC	Sets	L.R.	AIC	L.R.	AIC	L.R.	AIC
1	299.4	289.4	133.7	121.8	113.9	106.0	21	23.3	14.0	30.2	18.7	14.2	6.7
2	121.0	111.3	51.6	39.8	5.0	−2.8	22	26.0	16.3	25.2	13.6	13.0	5.2
3	165.9	156.0	49.3	37.4	14.2	6.2	23	35.5	25.9	29.1	17.7	18.5	10.7
4	78.5	68.8	27.2	15.5	24.0	16.1	24	34.2	24.6	17.6	6.1	10.9	3.2
5	137.9	128.0	45.7	33.8	32.8	24.9	25	37.7	28.4	30.2	19.1	16.0	8.5
6	78.2	68.4	30.7	19.0	4.7	−3.2	26	78.1	68.8	40.6	29.3	24.9	17.3
7	211.1	201.3	59.1	47.2	11.0	3.1	27	54.4	46.0	31.7	21.2	16.9	9.9
8	100.1	90.4	11.7	0.0	10.7	2.8	28	24.4	14.8	23.2	11.8	10.2	2.5
9	127.9	118.0	41.7	29.8	26.5	18.6	29	7.2	−2.1	9.6	−1.4	4.2	−3.3
10	163.3	153.5	91.3	79.6	29.5	21.6	30	69.7	60.0	63.2	51.8	44.2	36.5
11	50.0	40.1	35.6	23.9	9.1	1.2	31	29.6	20.4	22.5	11.5	10.5	3.1
12	33.9	24.6	25.2	14.1	23.4	16.0	32	121.4	111.8	30.2	18.7	11.1	3.3
13	220.7	210.8	30.5	18.6	14.0	6.1	33	59.8	50.5	33.2	22.1	9.9	2.3
14	103.5	93.7	58.8	47.1	8.7	0.8	34	31.7	22.0	36.6	25.1	24.1	16.3
15	41.2	31.5	23.0	11.4	7.8	0.0	35	53.8	44.8	65.8	53.8	16.6	9.1
16	46.0	36.6	33.2	22.0	12.7	5.0	36	34.3	24.6	23.6	12.0	13.2	5.4
17	80.3	70.5	53.8	42.1	17.6	9.7	37	12.6	3.4	12.9	2.2	0.7	−6.7
18	20.1	10.4	20.1	8.6	11.0	3.3	38	27.3	18.2	13.5	2.5	6.1	−1.0
19	47.4	37.7	37.7	26.2	9.8	2.0	39	27.3	18.1	6.0	−5.4	2.8	−4.8
20	54.6	44.8	43.1	31.4	25.9	18.0	40	4.8	−4.0	12.8	2.6	1.9	−4.9

Table 9. Accuracies and AUC for different logistic, autologistic, and independent logistic models.

Data	Logistic			Data	Logistic			Independent Logistic	
Sets	Modrel 1	Model 2	Auto	Sets	Model 1	Model 2	Auto	Model 1	Auto
Accuracy									
1	0.699	0.710	0.764	21	0.732	0.764	0.701	0.699	0.764
2	0.661	0.715	0.816	22	0.835	0.815	0.840	0.676	0.810
3	0.819	0.826	0.828	23	0.716	0.747	0.772	0.689	0.799
4	0.688	0.691	0.750	24	0.904	0.907	0.911	0.636	0.803
5	0.802	0.803	0.820	25	0.657	0.727	0.727	0.648	0.787
6	0.683	0.680	0.773	26	0.910	0.923	0.929	0.736	0.830
7	0.740	0.726	0.799	27	0.699	0.771	0.867	0.584	0.754
8	0.690	0.690	0.803	28	0.866	0.863	0.863	0.368	0.769
9	0.727	0.753	0.775	29	0.744	0.756	0.744	0.594	0.677
10	0.658	0.739	0.824	30	0.895	0.891	0.898	0.733	0.738
11	0.885	0.887	0.885	31	0.710	0.729	0.776	0.376	0.448
12	0.812	0.818	0.824	32	0.838	0.819	0.924	0.251	0.260
13	0.612	0.654	0.774	33	0.616	0.795	0.804	0.208	0.476
14	0.570	0.704	0.773	34	0.879	0.889	0.868	0.225	0.707
15	0.812	0.833	0.826	35	0.683	0.861	0.812	0.260	0.460
16	0.841	0.846	0.872	36	0.849	0.853	0.867	0.286	0.643
17	0.590	0.696	0.717	37	0.787	0.775	0.820	0.261	0.719
18	0.875	0.875	0.878	38	0.852	0.857	0.857	0.445	0.698
19	0.587	0.726	0.731	39	0.714	0.723	0.804	0.308	0.608
20	0.768	0.754	0.759	40	0.908	0.908	0.901	0.325	0.521
AUC									
1	65.27	70.06	81.63	21	70.22	80.71	75.68	61.12	70.09
2	56.52	72.21	79.97	22	63.23	70.48	75.22	52.64	56.91
3	60.08	68.37	86.12	23	65.33	70.46	72.66	49.05	50.09
4	64.53	65.86	79.01	24	64.84	74.15	80.73	50.38	48.04
5	64.96	67.50	82.86	25	66.98	81.55	83.06	53.03	51.56
6	57.72	72.21	80.66	26	68.42	74.93	85.79	50.73	50.81
7	54.49	70.62	85.98	27	71.72	83.79	95.03	49.22	52.01
8	60.72	61.09	83.06	28	61.79	75.30	75.00	47.54	44.78
9	64.07	68.69	81.68	29	65.54	71.32	66.52	48.28	49.88
10	68.98	80.16	87.32	30	72.33	83.30	82.66	49.37	49.61
11	59.39	72.28	83.57	31	67.53	76.27	82.13	46.17	46.65
12	77.50	77.83	79.40	32	63.84	81.33	91.61	47.56	46.91
13	58.39	62.03	84.52	33	61.31	77.19	88.58	48.62	48.99
14	57.18	75.98	82.99	34	69.72	71.82	79.49	47.79	51.57
15	60.07	64.84	79.30	35	70.71	91.96	88.49	50.41	51.07
16	73.37	81.94	85.37	36	63.52	73.58	75.74	49.22	47.93
17	60.91	71.91	77.24	37	55.38	78.42	71.58	53.92	49.57
18	61.60	74.04	68.29	38	67.07	64.38	73.90	47.94	44.55
19	56.66	75.76	77.97	39	56.89	62.27	76.91	48.67	48.72
20	62.76	74.54	75.98	40	57.60	76.50	71.63	50.77	51.20

The coefficient estimates, standard errors, and p-values for all autologistic models across all datasets are presented in Tables A1–A4 in the Appendix A. These tables present the estimated regression coefficients, standard errors, and p-Values (Prob). We also presented the same results for independent logistic regression models in Tables A5 and A6. The parameter estimates, standard errors, and p-values for the first model from Table A1 and the first model from Table A1 are the same, as both are the same Marginal models. However, if we compare Model 2 ($0 \rightarrow 1$) and Model 3 ($1 \rightarrow 1$) from Table A1 with the independent model and the independent logistic regression models in Table A5, we observe changes in the estimates, standard errors, and p-values. For example, the standard errors for the coefficient of elevation and area for Model 2 ($0 \rightarrow 1$) are 0.002 and 0.2927 with p-values of 0.0126 and 0.1892; compared with these, from the independent model for 2001 from Table A1, are 0.3000 and 0.0002 with p-values of 0.0487 and 0.0000, respectively. Now, while making inferences, we will accept the null hypothesis related to the elevation ($p \geq 0.05$) and will reject the null hypothesis related to the area ($p < 0.05$) for the independent model in Table A5. On the other hand, these inferences will reverse completely in the case of Model 2 ($0 \rightarrow 1$) from Table A1. Because we will reject the null hypothesis related to the elevation ($p < 0.05$) and will accept the null hypothesis related to the area ($p \geq 0.05$), respectively.

Table 10. Direction and significance level of coefficients for Markov regression models.

Variables/Models	**1**		**2**		**3**		**4**		**5**		**6**		**7**		**8**		**9**		**10**	
	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**
Intercept	–	**	–	**	–	**	–	**	–	**	–	**	–	**	–	**	-	**	–	**
Event	+	**	–		–		+		+	*	+		–		–		–		–	
Elevation	–	*	+	*	+		–		-		+		–		–		–		+	
Area	+	**	+		+	*	+		+	**	+		+	**	+		+	**	+	
Land	+		–		+		–		+		–		+		+		–		–	
AutoCov	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**
	11		**12**		**13**		**14**		**15**		**16**		**17**		**18**		**19**		**20**	
Intercept	–	**	–	**	–	**	–	**	–	**	–	**	–	**	–	**	–	**	–	**
Event	–		–	*	+		+		–		+		+		+	**	+	**	+	**
Elevation	–		–		–		–		+		+	*	+		–		–		–	
Area	+	*	+		+	*	+		+	*	–		+		+		+		+	*
Land	–		+		–		+		+		+		–		–		+		+	
AutoCov	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**
	21		**22**		**23**		**24**		**25**		**26**		**27**		**28**		**29**		**30**	
Intercept	–		–	**	–	*	–	**	–		–	**	–		–	**	-	**	-	**
Event	+		+	**	+		+	**	+	*	+	**	+		+	*	–	*	+	**
Elevation	+		–		+		+		–		+		+		–		+		+	
Area	–		–		-		+		–	*	+		–		+		–		–	
Land	–	**	+		+		–		–		–		–		–		+		+	
AutoCov	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**
	31		**32**		**33**		**34**		**35**		**36**		**37**		**38**		**39**		**40**	
Intercept	–		–		–	*	–	**	–		–	**	–		–	**	–	*	-	**
Event	+	*	+		+		+	**	+		+		–		–		+		–	
Elevation	+		+		–		+	*	–		+	*	–		–		+		+	
Area	–		–		–	*	–		–		–		–		–		–		+	
Land	–		–		–		+		–		+		–		–		–		–	
AutoCov	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**

** Significant at 1% level; * Significant at 5% level.

Table 11. Direction and significance level of coefficients for independent logistic regression models.

Variables/Models	**2000**		**2001**		**2002**		**2003**		**2004**		**2005**		**2006**		**2007**		**2008**		**2009**	
	$\hat{\beta}$	p.v	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**	$\hat{\beta}$	**p.v**
Intercept	−	**	−	**	−	**	−	**	−	**	−	**	−	**	−	**	−	**	−	**
Event	+	**	+		+		+		−		−		−		+		+	*	+	**
Elevation	−	*	+	*	−		+		−		−		-		+		+		−	
Area	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	*	+	
Land	+		+		+		+		+		−		−		+		+		+	
AutoCov	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**
	2010		2011		2012		2013		2014		2015		2016		2017		2018		2019	
Intercept	−	**	−	**	−	**	−	**	−	**	−	**	−	**	−	**	−	**	−	**
Event	+	**	+	**	+	**	+		+		+	**	+	*	+	**	+		+	
Elevation	+		+		−		−		+		+		+		+		+		+	
Area	+		−		−		−		+		−		−	*	−	**	−	**	−	*
Land	−		+		−		−		−		−		−		−		−		−	
AutoCov	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**	+	**

** Significant at 1% level; * Significant at 5% level.

5.3. Trajectory Risks Prediction

To compare the estimated trajectory risk, we used two stations as shown in Appendix A, Table A7. The observed values of the dependent variable (Y) and independent variables X_1 to X_4 for years 2000 to 2020 are also shown in Appendix A, Table A7. For Station 1, the Y for all the years was 1 ($PM_{2.5} \geq 35$ μg/m^3). However, for Station 2, the Y values were 0 $PM_{2.5} < 35$ μg/m^3 for all the years except for the year 2020. For Stations 1 and 2, we estimated the conditional and joint probabilities from the covariate values, as shown in Table A7. The trajectory risk for both stations is predicted using all four models (i–iv). The predicted conditional and joint probabilities from four models, for both stations, are shown in Table A8. The trajectory of conditional probabilities based on all four models for Station 1 is shown in Figure 3. Note that the conditional probability is based on the reduced sample, using the observed value of the conditioning variable Y (1 or 0). As all the values of Y are 1 for all the years, we expect that the predicted risk by the best model for each year should be as close as possible to 1. The predicted risk for this graph is the highest for most years from the autologistic regression model, followed by Model 1, Model 2, and Model 4, respectively. It is noteworthy that Model 4's prediction is the worst, as this is an independent model, i.e., without temporal dependence adjustment. Figure 4 displays the joint probability trajectory for Station 1 across all four models. This shows a pattern similar to that in Figure 3.

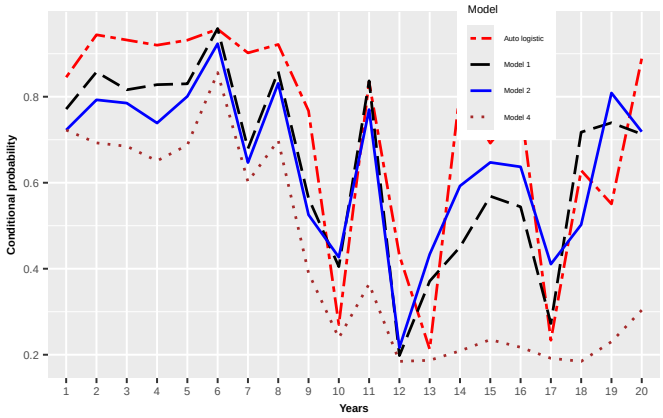


Figure 3. Trajectory of conditional probabilities for Station 1 using different models.

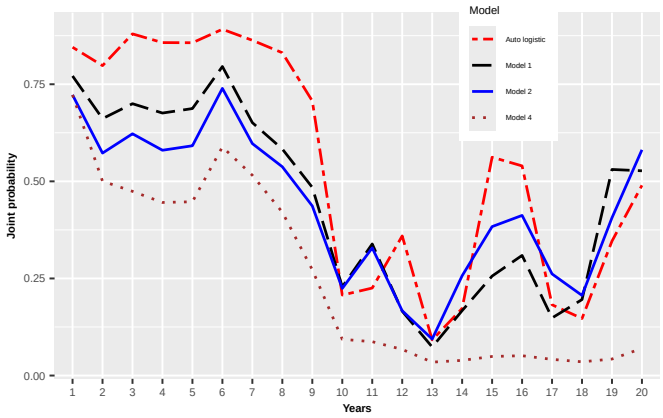


Figure 4. Trajectory of joint probabilities for Station 1 using different models.

The trajectory of conditional probabilities based on all four models for Station 2 is shown in Figure 5. Because all Y values are 0 for all years except 2020, we expect the best model's predicted risk for each year to be as close to 0 as possible. The model-predicted

risk for this graph is the lowest for most of the years from the autologistic regression model, followed by Model 1, Model 2, and Model 4, respectively. Figure 6 displays the joint probability trajectory for Station 2 across all four models. This shows a pattern similar to that in Figure 3.

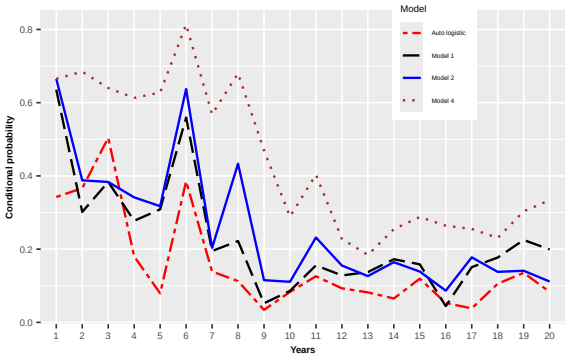


Figure 5. Trajectory of conditional probabilities for Station 2 using different models.

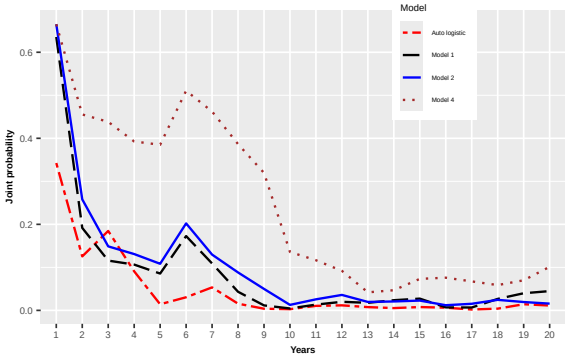


Figure 6. Trajectory of joint probabilities for Station 2 using different models.

We can generate new cases with varying covariate values and predict risk using the fitted models. Then, plotting these will help us understand the impact of distinct covariate categories on the trajectory risk. If we look at the covariate values for various years for Station 1 and Station 2 from Table A7, the values of Event, Area, and Land are the same for both years. However, Elevation is 0 for Station 1 and 353.84 for Station 2; as a result, the predicted risks for all the years are much lower for Station 2 as compared to Station 1 (Table A8).

6. Conclusions

In this paper, we propose a modeling approach and trajectory risk prediction for repeated binary outcomes from repeated-measures or longitudinal studies. The proposed framework for trajectory risk prediction employs a Markov chain to estimate joint probabilities as the trajectory risk. These joint probabilities are obtained by linking marginal and conditional probabilities from repeated-measures data. For many repeated-measures datasets from various fields, there is an inherent temporal and spatial dependence. The presence of temporal dependence may affect the standard error estimates of regression coefficients, and inferences will be misleading. Ultimately, it will affect the trajectory risk prediction. A Markov chain has a built-in mechanism that captures temporal dependencies. To measure the impact of covariates on the binary outcomes, we can adapt the logistic regression. The second problem is how we will adjust the spatial dependency simultaneously. In our proposed framework, we used autologistic regression to measure the covariate impact on binary outcomes. The autologistic regression adds a covariate: a spatially lagged term that accounts for spatial autocorrelation and is derived from latitude and longitude.

Using autologistics in the Markov chain framework, we adjust for both temporal and spatial dependencies.

The proposed framework is mathematically sound, and model fitting and trajectory prediction are straightforward and computationally efficient. We illustrated the proposed framework using real-life longitudinal data on fine-particle ($PM_{2.5}$) outdoor air pollution from the United States between the years 2000 and 2020. We compared four models of trajectory risk, as described in the result sections. The highest prediction accuracy was achieved using autologistic regression within the proposed Markov chain framework, as it accounts for both temporal and spatial dependences. The second-best model was logistic regression, including latitude and longitude as covariates along with other available independent variables. However, this model does not adjust for temporal dependency. The third model uses the available independent variables, excluding latitude and longitude, as covariates. The fourth model, which assumed no temporal or spatial dependencies, produced the lowest prediction accuracy.

One can quickly implement the proposed framework using available packages in R or other software, e.g., Python, to fit an auto logistic regression. The augmented data preparation structure simplifies computation to univariate cases, and parallel computing enables us to fit all models simultaneously. It is straightforward to include a varying number of covariates at each follow-up, as many new factors may be added in later follow-ups, as in many longitudinal studies. The trajectory risk can be predicted for a segment or up to the endpoint, both conditionally and jointly. The proposed framework can be readily applied to analyze datasets from various fields that exhibit inherent temporal and spatial dependencies.

7. Limitations

Our objectives were to develop a modeling framework to simultaneously adjust for both temporal and spatial dependence in longitudinal datasets and to test hypotheses and perform statistical inference. Hence, we used statistical modeling approaches. However, in our proposed framework, we can use any machine-learning (ML) model for a cross-sectional dataset to predict the trajectory risk. But it should be noted that ML models lack statistical hypothesis testing and inference. This is possible because our proposed framework reduces a multivariate problem into univariate cases. It is also possible to use a Long Short-Term Memory (LSTM) model if only the prediction is of interest. However, this is beyond the scope of this paper, as ML models are not of interest to us. Nevertheless, it will be interesting work, which we plan to pursue in the future. Also, it would be interesting to check the data split into training and test sets for model validation, as we do in the ML approach.

Author Contributions: Conceptualization: R.C. and M.T.H.; methodology: R.C.; software, R.C.; validation, R.C. and M.T.H.; formal analysis, R.C.; resources: R.C. and M.T.H.; data curation, R.C.; writing—review and editing: R.C. and M.T.H.; visualization: R.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2017-04246) and Atlantic Association for Research in the Mathematical Sciences (AARMS).

Informed Consent Statement: These longitudinal datasets are collected by the United States Environmental Protection Agency (EPA) and are public domain.

Data Availability Statement: These longitudinal datasets are obtained from the United States Environmental Protection Agency (EPA) for making these data publicly available. The datasets are publicly available on the EPA website. The processed data and the R codes for model fitting will be made available from the researchgate site: https://www.researchgate.net/profile/Rafiqul-Chowdhury-2 (accessed on 10 February 2026).

Acknowledgments: The authors also acknowledge the data from the HRS (Health and Retirement Study) that is sponsored by the National Institute of Aging and conducted by the University of Michigan. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2017-04246) and Atlantic Association for Research in the Mathematical Sciences (AARMS). We also acknowledge the United States Environmental Protection Agency (EPA) for making these data publicly available.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

$PM_{2.5}$	Fine particulate matter
AUC	Area under the curve
ROC	Receiver operating Ccharacteristic
ML	Machine learning
LSTM	Long short-term memory

Appendix A

Table A1. Parameter estimates of Markov regression (autologistic) models.

Variable	Coef	StdError	Prob	Coef	StdError	Prob
	Marginal Model 1			Model 2 (0 → 1)		
Intercept	−6.2864	0.5500	0.0000	−5.0212	0.5929	0.0000
Event	1.0963	0.3166	0.0005	−0.1163	0.5585	0.8350
Elevation	−0.0003	0.0001	0.0243	0.0005	0.0002	0.0126
Area	1.2246	0.1831	0.0000	0.3843	0.2927	0.1892
Land	0.1923	0.1478	0.1931	−0.0919	0.2679	0.7316
AutoCov	8.5286	0.7274	0.0000	10.8586	1.2969	0.0000
	Model 3 (1 → 1)			Model 4 (0 → 1)		
Intercept	−10.2453	1.0629	0.0000	−3.6229	0.5854	0.0000
Event	−0.1960	0.3708	0.5970	0.3025	0.3789	0.4247
Elevation	0.0004	0.0004	0.3173	−0.0004	0.0004	0.3062
Area	0.7777	0.3144	0.0134	0.3139	0.3071	0.3066
Land	0.3716	0.2273	0.1020	−0.4836	0.2679	0.0711
AutoCov	13.1165	1.2544	0.0000	7.9665	1.2462	0.0000
	Model 5 (1 → 1)			Model 6 (0 → 1)		
Intercept	−10.0898	1.2626	0.0000	−4.1620	0.5738	0.0000
Event	0.6634	0.3149	0.0352	0.4718	0.4061	0.2453
Elevation	−0.0002	0.0002	0.3010	0.0003	0.0002	0.2918
Area	0.9193	0.2973	0.0020	0.3148	0.3356	0.3482
Land	0.2761	0.2190	0.2074	−0.0117	0.2808	0.9669
AutoCov	12.7512	1.4763	0.0000	8.8053	1.1943	0.0000
	Model 7 (1 → 1)			Model 8 (0 → 1)		
Intercept	−8.0056	0.7877	0.0000	−4.6570	0.6535	0.0000
Event	−0.2201	0.5113	0.6669	−0.4298	0.3962	0.2780
Elevation	−0.0000	0.0002	0.8018	−0.0004	0.0003	0.1961
Area	0.8153	0.3037	0.0073	0.3745	0.3535	0.2895
Land	0.2867	0.2199	0.1923	0.1614	0.2977	0.5878
AutoCov	10.9660	0.9451	0.0000	10.7189	1.4810	0.0000
	Model 9 (1 → 1)			Model 10 (0 → 1)		
Intercept	−7.2458	0.9089	0.0000	−4.8653	0.7027	0.0000
Event	−0.0708	0.3750	0.8503	−0.5473	0.4114	0.1834
Elevation	−0.0003	0.0003	0.2306	0.0001	0.0002	0.6256
Area	1.1449	0.3011	0.0001	0.1567	0.3342	0.6392
Land	−0.1363	0.2212	0.5377	−0.5383	0.2965	0.0695
AutoCov	9.9790	1.1427	0.0000	9.3056	1.0175	0.0000

Table A2. Parameter estimates of Markov regression (autologistic) models.

Variable	Coef	StdError	Prob	Coef	StdError	Prob
		Model 11 (1 → 1)			Model 12 (0 → 1)	
Intercept	−5.6488	1.2845	0.0000	−2.1864	0.6914	0.0016
Event	−0.6411	0.5085	0.2074	−2.0117	0.8551	0.0186
Elevation	−0.0003	0.0003	0.4117	−0.0008	0.0005	0.1263
Area	0.8791	0.4057	0.0302	0.1481	0.4875	0.7613
Land	−0.2206	0.3150	0.4838	0.3433	0.4604	0.4558
AutoCov	8.1059	1.3775	0.0000	4.1056	1.3957	0.0033
		Model 13 (1 → 1)			Model 14 (0 → 1)	
Intercept	−6.4196	0.6335	0.0000	−5.0657	0.7048	0.0000
Event	0.2506	0.3775	0.5068	0.4112	0.2916	0.1586
Elevation	−0.0002	0.0002	0.4804	−0.0005	0.0005	0.3146
Area	0.7202	0.3063	0.0187	0.3098	0.3440	0.3678
Land	−0.3570	0.2096	0.0885	0.0559	0.2847	0.8443
AutoCov	10.6605	0.9455	0.0000	8.7850	1.1128	0.0000
		Model 15 (1 → 1)			Model 16 (0 → 1)	
Intercept	−7.6489	1.6367	0.0000	−3.9948	0.6971	0.0000
Event	−0.2083	0.3654	0.5687	0.9551	0.5137	0.0630
Elevation	0.0000	0.0004	0.9318	0.0010	0.0005	0.0532
Area	1.1207	0.4450	0.0118	−0.7320	0.5161	0.1561
Land	0.1229	0.3330	0.7121	0.2723	0.4613	0.5549
AutoCov	9.7708	1.8675	0.0000	12.7115	3.0904	0.0000
		Model 17 (1 → 1)			Model 18 (0 → 1)	
Intercept	−5.1393	0.7861	0.0000	−3.0158	0.5948	0.0000
Event	0.7931	0.4329	0.0669	1.5026	0.4437	0.0007
Elevation	0.0002	0.0005	0.6968	−0.0005	0.0007	0.5049
Area	0.8677	0.3899	0.0260	0.2049	0.4812	0.6702
Land	−0.0389	0.2466	0.8748	−0.1517	0.3651	0.6777
AutoCov	7.8228	1.1619	0.0000	5.8695	1.9056	0.0021
		Model 19 (1 → 1)			Model 20 (0 → 1)	
Intercept	−3.4099	0.7558	0.0000	−3.7591	0.5567	0.0000
Event	1.8647	0.6710	0.0055	1.3833	0.4871	0.0045
Elevation	−0.0004	0.0005	0.3958	−0.0001	0.0006	0.8644
Area	0.4447	0.5060	0.3795	0.9549	0.4277	0.0256
Land	0.1335	0.3125	0.6693	0.1605	0.2733	0.5570
AutoCov	5.7283	1.0588	0.0000	5.6624	1.1408	0.0000

Table A3. Parameter estimates of Markov regression (autologistic) models.

Variable	Coef	StdError	Prob	Coef	StdError	Prob
		Model 21 (1 → 1)			Model 22 (0 → 1)	
Intercept	−1.8073	1.3453	0.1791	−2.3284	0.5576	0.0000
Event	1.6870	0.9314	0.0701	1.9788	0.5471	0.0003
Elevation	0.0004	0.0007	0.5941	−0.0004	0.0006	0.5282
Area	−0.2881	0.6869	0.6749	−0.3632	0.4214	0.3888
Land	−1.2841	0.4571	0.0050	0.0153	0.3730	0.9674
AutoCov	4.6732	1.6619	0.0049	5.0858	1.4814	0.0006
		Model 23 (1 → 1)			Model 24 (0 → 1)	
Intercept	−1.7299	0.7152	0.0156	−3.8102	0.7903	0.0000
Event	1.0363	0.6105	0.0896	2.6031	0.9344	0.0053
Elevation	0.0001	0.0006	0.8697	0.0007	0.0007	0.3705
Area	−0.8869	0.5720	0.1210	0.4262	0.7387	0.5640
Land	0.2952	0.3715	0.4269	−0.6226	0.4564	0.1725
AutoCov	4.7579	1.3219	0.0003	9.7584	2.1118	0.0000
		Model 25 (1 → 1)			Model 26 (0 → 1)	
Intercept	−1.7922	1.0107	0.0762	−4.5747	0.9890	0.0000
Event	1.8583	0.8867	0.0361	2.1535	1.3225	0.1035
Elevation	−0.0003	0.0006	0.6383	0.0001	0.0007	0.9005
Area	−1.3821	0.6656	0.0379	0.9675	0.9440	0.3054
Land	−0.0584	0.4966	0.9064	−0.6020	0.5200	0.2470
AutoCov	5.8954	1.4691	0.0001	13.0182	2.2611	0.0000

Table A3. *Cont.*

Variable	Coef	StdError	Prob	Coef	StdError	Prob
		Model 27 (1 → 1)			Model 28 (0 → 1)	
Intercept	−2.2883	2.3872	0.3378	−3.7804	0.7710	0.0000
Event	0.6607	1.2306	0.5913	1.0336	0.5244	0.0487
Elevation	0.0019	0.0020	0.3507	−0.0006	0.0008	0.4682
Area	−2.8265	1.9133	0.1396	1.2037	0.7108	0.0904
Land	−0.5845	0.6994	0.4033	−0.2222	0.3536	0.5297
AutoCov	8.8956	2.0731	0.0000	6.1995	1.6428	0.0002
		Model 29 (1 → 1)			Model 30 (0 → 1)	
Intercept	−0.6390	1.0604	0.5468	−3.5876	0.6256	0.0000
Event	−0.1290	0.7431	0.8622	2.2743	0.5496	0.0000
Elevation	0.0007	0.0007	0.2923	0.0004	0.0005	0.3831
Area	−0.2034	0.7186	0.7772	−0.2344	0.5432	0.6661
Land	0.4877	0.5255	0.3534	0.0476	0.4147	0.9087
AutoCov	2.1329	1.2516	0.0884	8.5465	1.7587	0.0000

Table A4. Parameter estimates of Markov regression (autologistic) models.

Variable	Coef	StdError	Prob	Coef	StdError	Prob
		Model 31 (1 → 1)			Model 32 (0 → 1)	
Intercept	−2.5904	1.3901	0.0624	−4.8253	0.7228	0.0000
Event	2.1922	0.9768	0.0248	0.4852	1.0939	0.6574
Elevation	0.0000	0.0007	0.9499	0.0014	0.0008	0.0674
Area	−1.3636	0.8210	0.0968	−0.5521	0.5791	0.3404
Land	−0.3239	0.4955	0.5134	−0.8674	0.4900	0.0767
AutoCov	6.4315	1.7133	0.0002	15.6361	2.0414	0.0000
		Model 33 (1 → 1)			Model 34 (0 → 1)	
Intercept	−1.6897	0.8372	0.0436	−2.8245	0.6114	0.0000
Event	1.0683	0.8129	0.1887	2.2958	0.7876	0.0036
Elevation	−0.0005	0.0006	0.4306	0.0017	0.0008	0.0279
Area	−1.2687	0.6340	0.0454	−0.5252	0.4884	0.2822
Land	−0.3603	0.5014	0.4725	0.0675	0.3887	0.8622
AutoCov	6.5321	1.2128	0.0000	6.4246	2.3407	0.0061
		Model 35 (1 → 1)			Model 36 (0 → 1)	
Intercept	−1.2678	0.7492	0.0906	−2.6542	0.5850	0.0000
Event	3.1071	2.0486	0.1293	1.5773	1.1951	0.1869
Elevation	−0.0003	0.0006	0.5689	0.0012	0.0006	0.0454
Area	−0.9603	0.6189	0.1207	−0.8400	0.4496	0.0617
Land	−0.8640	0.5319	0.1043	0.0792	0.3671	0.8293
AutoCov	5.5113	1.0867	0.0000	8.4485	2.0549	0.0000
		Model 37 (1 → 1)			Model 38 (0 → 1)	
Intercept	−2.0199	1.4773	0.1715	−3.1919	0.7853	0.0000
Event	−0.0948	0.9156	0.9176	−2.5685	4.8236	0.5944
Elevation	−0.0007	0.0006	0.2310	−0.0010	0.0011	0.3591
Area	−0.3723	0.6472	0.5651	−0.1842	0.5448	0.7353
Land	−0.3015	0.5585	0.5893	−0.1842	0.3874	0.6344
AutoCov	5.4078	1.9127	0.0047	13.0873	3.0425	0.0000
		Model 39 (1 → 1)			Model 40 (0 → 1)	
Intercept	−3.1831	1.3085	0.0150	−3.3113	1.1089	0.0028
Event	1.2232	1.5812	0.4391	−1.8867	3.1949	0.5548
Elevation	0.0001	0.0004	0.7961	0.0007	0.0009	0.3988
Area	−0.3580	0.5710	0.5307	0.6003	1.0090	0.5519
Land	−0.3478	0.4845	0.4728	−0.2509	0.5839	0.6674
AutoCov	6.6194	1.6389	0.0001	6.7866	4.0702	0.0954

Table A5. Parameter estimates of marginal and independent logistic regression models.

Variables/Models	Coef	StdError	Prob	Coef	StdError	Prob
Years		2000			2001	
Intercept	−6.2864	0.5500	0.0000	−9.3119	0.6159	0.0000
Event	1.0963	0.3166	0.0005	0.0442	0.3000	0.8830
Elevation	−0.0003	0.0001	0.0243	0.0003	0.0002	0.0487
Area	1.2246	0.1831	0.0000	0.9436	0.2075	0.0000
Land	0.1923	0.1478	0.1931	0.1294	0.1663	0.4364
AutoCov	8.5286	0.7274	0.0000	13.2335	0.8371	0.0000
		2002			2003	
Intercept	−9.0220	0.6832	0.0000	−6.9910	0.4978	0.0000
Event	0.1446	0.2418	0.5497	0.2048	0.3022	0.4980
Elevation	−0.0000	0.0001	0.9952	0.0001	0.0001	0.5466
Area	0.8903	0.2081	0.0000	0.7881	0.2188	0.0003
Land	0.0051	0.1653	0.9754	0.1854	0.1673	0.2678
AutoCov	13.0287	0.9568	0.0000	10.4660	0.6814	0.0000
		2004			2005	
Intercept	−7.9360	0.6279	0.0000	−7.9177	0.7289	0.0000
Event	−0.0015	0.2593	0.9953	−0.3377	0.3019	0.2634
Elevation	−0.0002	0.0002	0.3356	−0.0002	0.0002	0.2528
Area	0.9952	0.2323	0.0000	0.8728	0.2463	0.0004
Land	0.0493	0.1746	0.7777	−0.3370	0.2038	0.0982
AutoCov	12.2030	0.9197	0.0000	11.4128	0.8900	0.0000
		2006			2007	
Intercept	−6.3137	0.5310	0.0000	−7.7929	0.7637	0.0000
Event	−0.0947	0.2815	0.7365	0.2172	0.2204	0.3246
Elevation	−0.0002	0.0002	0.3683	0.0000	0.0003	0.9176
Area	0.7090	0.2530	0.0051	0.7964	0.2694	0.0031
Land	−0.2194	0.1824	0.2291	0.0186	0.2096	0.9294
AutoCov	11.1199	0.8633	0.0000	11.1521	1.0025	0.0000
		2008			2009	
Intercept	−5.8153	0.5676	0.0000	−4.1083	0.4718	0.0000
Event	0.7509	0.3088	0.0150	1.3477	0.3410	0.0001
Elevation	0.0004	0.0003	0.2757	−0.0001	0.0004	0.7185
Area	0.6686	0.3030	0.0273	0.4501	0.3264	0.1679
Land	0.0667	0.2123	0.7533	0.0092	0.2296	0.9681
AutoCov	10.4392	1.0119	0.0000	8.6782	1.0340	0.0000

Table A6. Parameter estimates of independent logistic regression models.

	Coef	StdError	Prob	Coef	StdError	Prob
Years		2010			2011	
Intercept	−4.5469	0.5194	0.0000	−2.3833	0.4087	0.0000
Event	1.5623	0.4292	0.0003	1.4747	0.4021	0.0002
Elevation	0.0001	0.0004	0.7421	0.0001	0.0004	0.8479
Area	0.6076	0.3204	0.0579	−0.3224	0.3153	0.3065
Land	−0.2453	0.2226	0.2703	0.1129	0.2478	0.6488
AutoCov	8.3403	0.9802	0.0000	5.2163	1.0611	0.0000
		2012			2013	
Intercept	−2.4737	0.4390	0.0000	−3.1052	0.4719	0.0000
Event	2.1831	0.5476	0.0001	1.3306	0.7726	0.0850
Elevation	−0.0003	0.0004	0.4382	−0.0000	0.0004	0.9497
Area	−0.4317	0.3544	0.2232	−0.5571	0.3706	0.1328
Land	−0.0202	0.2774	0.9419	−0.3123	0.3119	0.3167
AutoCov	6.4167	1.1568	0.0000	10.3072	1.2599	0.0000
		2014			2015	
Intercept	−2.9508	0.4357	0.0000	−2.9107	0.4348	0.0000
Event	0.6063	0.4065	0.1358	1.7506	0.4433	0.0001
Elevation	0.0002	0.0003	0.5472	0.0003	0.0003	0.4429
Area	0.0025	0.3458	0.9943	−0.5745	0.3405	0.0915
Land	−0.1694	0.2532	0.5034	−0.1868	0.2699	0.4889
AutoCov	7.0504	1.0190	0.0000	7.5818	1.0359	0.0000

Table A6. *Cont.*

	Coef	StdError	Prob	Coef	StdError	Prob
Years		2016			2017	
Intercept	−3.5205	0.4815	0.0000	−2.0276	0.4115	0.0000
Event	1.3694	0.6385	0.0320	2.2611	0.6432	0.0004
Elevation	0.0006	0.0004	0.1508	0.0004	0.0003	0.2865
Area	−0.8632	0.3483	0.0132	−0.9598	0.3298	0.0036
Land	−0.2037	0.2919	0.4852	−0.3181	0.2776	0.2518
AutoCov	11.1475	1.3242	0.0000	6.8335	1.0986	0.0000
		2018			2019	
Intercept	−2.1919	0.4629	0.0000	−2.6640	0.5526	0.0000
Event	0.8771	0.7341	0.2322	0.8847	0.7789	0.2561
Elevation	0.0007	0.0004	0.0538	0.0003	0.0003	0.3418
Area	−1.0251	0.3208	0.0014	−0.6779	0.3152	0.0315
Land	−0.0761	0.2608	0.7704	−0.1547	0.2540	0.5425
AutoCov	7.3146	1.1536	0.0000	8.2038	1.3921	0.0000

Table A7. Values for dependent and independent variables for two selected monitoring stations.

Station 1					Station 2				
Y	**Event**	**Elevation**	**Area**	**Land**	**Y**	**Event**	**Elevation**	**Area**	**Land**
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	0	0	353.84	1	0
1	0	0	1	0	1	0	353.84	1	0

Table A8. Predicted marginal, conditional, and joint risks for two selected stations.

		Model 1		Model 2		Model 4		Auto Logistic	
Year	**Y**	**CP**	**JP**	**CP**	**JP**	**CP**	**JP**	**CP**	**JP**
					Station 1				
1	1	0.771	0.771	0.723	0.723	0.723	0.723	0.845	0.845
2	1	0.858	0.661	0.793	0.573	0.693	0.500	0.944	0.798
3	1	0.816	0.700	0.785	0.622	0.685	0.474	0.932	0.880
4	1	0.828	0.676	0.739	0.580	0.651	0.446	0.920	0.857
5	1	0.830	0.687	0.801	0.592	0.687	0.447	0.932	0.857
6	1	0.959	0.796	0.923	0.739	0.856	0.589	0.957	0.892
7	1	0.679	0.651	0.647	0.597	0.602	0.516	0.902	0.863
8	1	0.859	0.583	0.831	0.538	0.698	0.421	0.921	0.831
9	1	0.564	0.484	0.525	0.437	0.390	0.272	0.767	0.707
10	1	0.405	0.228	0.427	0.224	0.239	0.093	0.270	0.207
11	1	0.837	0.339	0.770	0.329	0.365	0.087	0.834	0.225
12	1	0.198	0.166	0.217	0.167	0.185	0.067	0.431	0.359
13	1	0.371	0.073	0.434	0.094	0.187	0.035	0.213	0.092
14	1	0.450	0.167	0.593	0.257	0.209	0.039	0.812	0.173
15	1	0.569	0.256	0.647	0.383	0.235	0.049	0.692	0.562
16	1	0.544	0.309	0.637	0.412	0.217	0.051	0.780	0.540
17	1	0.273	0.148	0.411	0.262	0.192	0.042	0.234	0.182
18	1	0.718	0.196	0.502	0.206	0.185	0.035	0.629	0.147
19	1	0.739	0.531	0.809	0.406	0.230	0.042	0.551	0.346
20	1	0.713	0.527	0.719	0.581	0.304	0.070	0.889	0.490

Table A8. *Cont.*

Year	Y	Model 1		Model 2		Model 4		Auto Logistic	
		CP	JP	CP	JP	CP	JP	CP	JP
					Station 2				
1	0	0.636	0.636	0.665	0.665	0.665	0.665	0.343	0.343
2	0	0.302	0.192	0.388	0.258	0.685	0.455	0.366	0.126
3	0	0.384	0.116	0.384	0.149	0.640	0.438	0.504	0.185
4	0	0.277	0.106	0.341	0.131	0.613	0.393	0.180	0.091
5	0	0.309	0.086	0.317	0.108	0.628	0.385	0.079	0.014
6	0	0.560	0.173	0.637	0.202	0.813	0.511	0.385	0.031
7	0	0.195	0.109	0.203	0.130	0.568	0.462	0.139	0.053
8	0	0.222	0.043	0.433	0.088	0.680	0.387	0.113	0.016
9	0	0.052	0.012	0.115	0.050	0.470	0.320	0.034	0.004
10	0	0.086	0.004	0.111	0.013	0.289	0.136	0.083	0.003
11	0	0.156	0.013	0.232	0.026	0.403	0.117	0.126	0.010
12	0	0.128	0.020	0.155	0.036	0.228	0.092	0.093	0.012
13	0	0.137	0.018	0.126	0.020	0.184	0.042	0.082	0.008
14	0	0.172	0.024	0.164	0.021	0.254	0.047	0.065	0.005
15	0	0.159	0.027	0.139	0.023	0.288	0.073	0.119	0.008
16	0	0.044	0.007	0.087	0.012	0.264	0.076	0.053	0.006
17	0	0.151	0.007	0.178	0.015	0.255	0.067	0.038	0.002
18	0	0.177	0.027	0.138	0.024	0.231	0.059	0.106	0.004
19	0	0.225	0.040	0.141	0.019	0.303	0.070	0.135	0.014
20	0	0.199	0.045	0.111	0.016	0.334	0.101	0.083	0.011

CP = Conditional Probability; JP = Joint Probability.

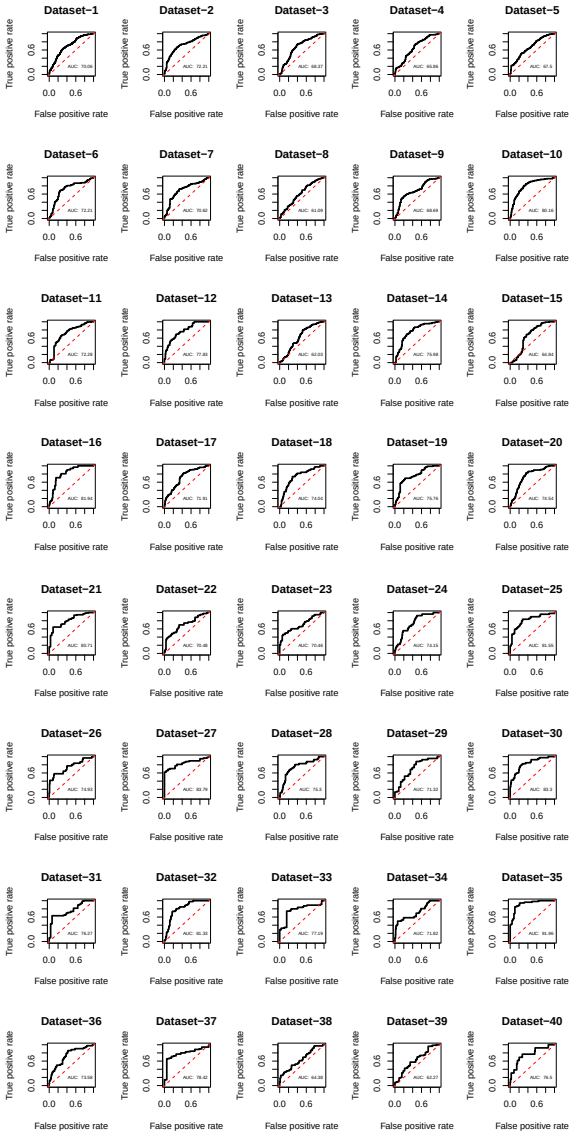


Figure A1. ROC using logistic regression Model 1 for 41 datasets.

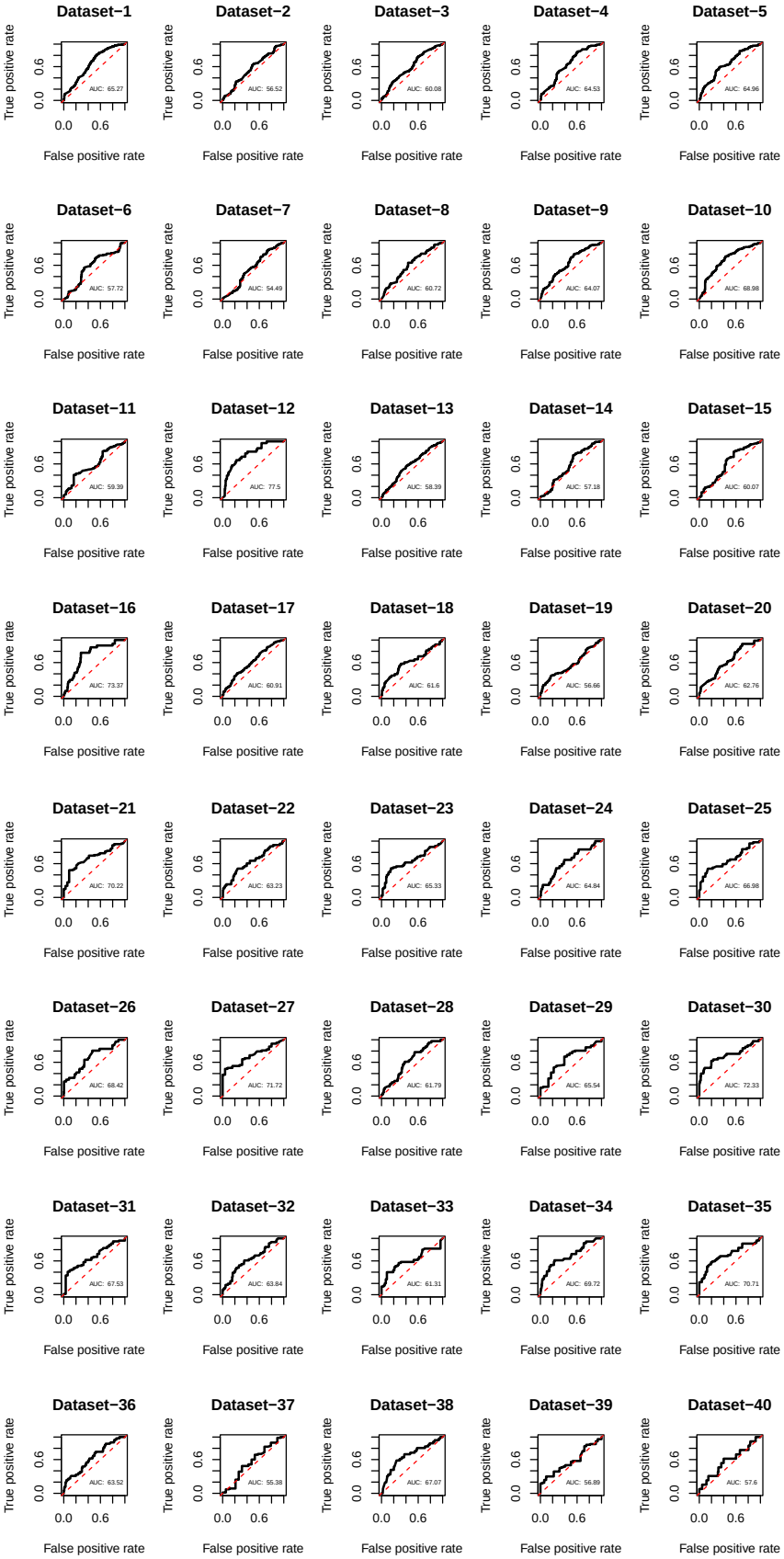


Figure A2. ROC using logistic regression Model 2 for 41 datasets.

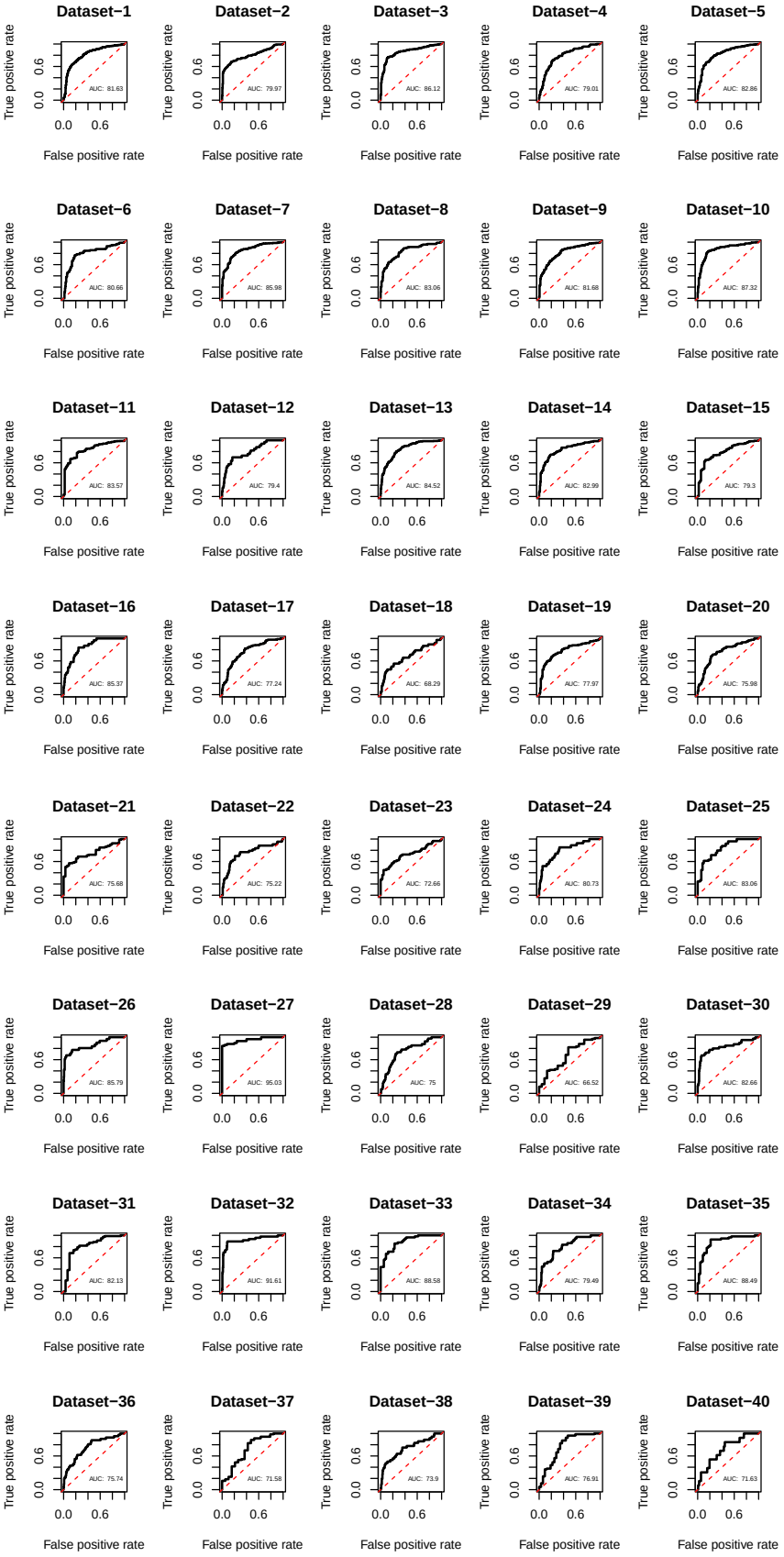


Figure A3. ROC using autologistic regression model for 41 datasets.

Table A9. R-codes for fitting autologistic and logistic regression models.

```
library(foreach);library(doParallel);library(tmap);library(spdep);
library(sf);library(ape);library(rms);library(spatialEco);library(psych)
require(sp);library(autoReg);

Dat41 <- Data41Sets[,c(4:9,12,16:17)]
indvars <- c("Event", "Elevation", "Area", "Land")
  idvar  <- "SiteID"
  depvar <- "Y"
  tvar   <- "Type1"
  b      <- 5
  totfol <- 40

  stopImplicitCluster()
  registerDoParallel(6)
  system.time(
  Auto40 <- foreach(i= 1:totfol,
 packages=c('MASS','sf','spdep','spatialEco','rms','dplyr','caret','ape'))
 %dopar% {
 newDat <- Data41Sets[Data41Sets$DivID==i,]
 newDat <- newDat[complete.cases(newDat),]
 colnames(newDat)[16]<-"x"
 colnames(newDat)[17]<-"y"
 newDat <- st_as_sf(newDat, coords = c("x", "y"), crs = 28992,
                                agr = "constant")
 fitmod <- logistic.regression(newDat, y='Y', x=c('Event', 'Elevation',
                'Area', 'Land'),
                autologistic=TRUE, coords=st_coordinates(newDat),
                bw=5000)
 aa   <- fitmod$model
 bb   <- fitmod$diagTable
 cc   <- fitmod$coefTable
 prob <- predict(aa, type='fitted.ind')
 Yh     <- ifelse(prob > 0.50, 1, 0)
 conFM <-    confusionMatrix(data=as.factor(Yh),
                    reference = as.factor(newDat$Y))
 prob   <- as.data.frame(cbind(SiteID=newDat$SiteID,
                                      Y=newDat$Y,Yh,prob))
 colnames(prob) <- c(idvar, paste("Y",i,sep=""), paste("Yh",i,sep=""),
                         paste("Py",i,sep=""))
        %%%%% Logistic model
         fitmod1 <- logistic.regression(newDat, y='Y', x=c('Event',
                          'Elevation','Area', 'Land'))
         aa1   <- fitmod1$model
         bb1   <- fitmod1$diagTable
         cc1   <- fitmod1$coefTable
         prob1 <- predict(aa1, type='fitted.ind')
         Yhh     <- ifelse(prob1 > 0.50, 1, 0)
 conFM1 <-    confusionMatrix(data=as.factor(Yhh),
                       reference = as.factor(newDat$Y))
 prob1   <- as.data.frame(cbind(SiteID=newDat$SiteID,
                                        Y=newDat$Y,Yh,prob1))
 colnames(prob1) <- c(idvar, paste("Y",i,sep=""), paste("Yh",i,sep=""),
                         paste("Py",i,sep=""))

         mls    <- list(fitmod,prob,conFM,fitmod1,prob1,conFM1)
           })
  stopImplicitCluster()
```

References

1. Murray, C.J.L.; Aravkin, A.Y.; Zheng, P.; Abbafati, C.; Abbas, K.M.; Abbasi-Kangevari, M.; Abd-Allah, F.; Abdelalim, A.; Abdollahi, M.; Abdollahpour, I.; et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *Lancet* **2019**, *396*, 1223–1249. [CrossRef]
2. Wu, W.; Zhang, Y. Effects of particulate matter ($PM_{2.5}$) and associated acidity on ecosystem functioning: Response of leaf litter breakdown. *Environ. Sci. Pollut. Res.* **2018**, *25*, 30720–30727. [CrossRef]
3. Environmental Protection Agency (EPA). United States, Pre-Generated Data Files. 2020. Available online: https://aqs.epa.gov/aqsweb/airdata//download_files.html (accessed on 10 January 2022).
4. Xu, F.; Shi, X.; Qiu, X.; Jiang, X.; Fang, Y.; Wang, J.; Hu, D.; Zhu, T. Investigation of the chemical components of ambient fine particulate matter ($PM_{2.5}$) associated with in vitro cellular responses to oxidative stress and inflammation. *Environ. Int.* **2020**, *136*, 105475. [CrossRef]
5. Jin, J.; Liu, S.; Wang, L.; Wu, S.; Zhao, W. Fractional Vegetation Cover and Spatiotemporal Variations of PM2.5 Concentrations in the Beijing Tianjin Hebei Region of China. *Atmosphere* **2022**, *13*, 1850. [CrossRef]
6. Ryu, J.; Kim, J.J.; Byeon, H.; Go, T.; Lee, S.J. Removal of fine particulate matter ($PM_{2.5}$) via atmospheric humidity caused by evapotranspiration. *Environ. Pollut.* **2019**, *245*, 253–259. [CrossRef]
7. Bilal, M.; Janet, E.; Nazeer, M.; Shi, Y.; Wang, L.; Raghavendra Kumar, K.; Ho, H.C.; Mazhar, U.; Bleiweiss, M.P.; Qiu, Z.; et al. Characteristics of Fine Particulate Matter ($PM_{2.5}$) over Urban, Suburban, and Rural Areas of Hong Kong. *Atmosphere* **2019**, *10*, 496. [CrossRef]
8. Chen, L.; Shi, L. Differences in urban–rural gradient and driving factors of $PM_{2.5}$ concentration in the Zhengzhou Metropolitan Area. *Air Qual. Atmos. Health* **2024**, *17*, 2187–2201. [CrossRef]
9. Tucker, W.G. Overview of $PM_{2.5}$ sources and control strategies. *Fuel Process. Technol.* **2000**, *65*, 379–392. [CrossRef]
10. Zhang, X.; Bao, Z.; Zhang, L.; Zhou, J.; Che, H.; Li, Q.; Tian, M.; Yang, F.; Chen, Y. Biomass burning and aqueous reactions drive the elevation of wintertime $PM_{2.5}$ in the rural area of the Sichuan basin, China. *Atmos. Environ.* **2023**, *306*, 119779. [CrossRef]
11. Huang, R.J.; Zhang, Y.; Bozzetti, C.; Ho, K.; Cao, J.; Han, Y.; Daellenbach, K.R.; Slowik, J.G.; Platt, S.M.; Canonaco, F.; et al. High secondary aerosol contribution to particulate pollution during haze events in China. *Nature* **2014**, *514*, 218–222. [CrossRef] [PubMed]
12. Chen, Z.; Chen, D.; Zhao, C.; Kwan, M.; Cai, J.; Zhuang, Y.; Zhao, B.; Wang, X.; Chen, B.; Yang, J.; et al. Influence of meteorological conditions on PM2.5 concentrations across China: A review of methodology and mechanism. *Environ. Int.* **2020**, *139*, 105558. [CrossRef]
13. Nadhira, D.; Latif, M.T.; Muda, K.; Hussein, N. Influence of Meteorological Variables on Suburban Atmospheric $PM_{2.5}$ of Southern Region of Peninsular Malaysia. *Aerosol Air Qual. Res.* **2020**, *20*, 14–25. [CrossRef]
14. Das, K.; Chatterjee, D.; Jana, D.; Bhattacharya, R.K. Application of land use regression model with regularization algorithm to asses PM25 and Pm10 concentration and health risk in Kolkata Metropolitan. *Urban Clim.* **2023**, *49*, 101473. [CrossRef]
15. Wang, R.; Henderson, S.B.; Sbihi, H.; Allen, R.W.; Brauer, M. Temporal stability of land use regression models for traffic-related air pollution. *Atmos. Environ.* **2013**, *64*, 312–319. [CrossRef]
16. Islam, M.A.; Chowdhury, R.I. A higher-order Markov model for analyzing covariate dependence. *Appl. Math. Model.* **2006**, *30*, 477–488. [CrossRef]
17. Islam, M.A.; Chowdhury, R.I.; Huda, S. *Markov Models with Covariate Dependence for Repeated Measures*; Nova Science: New York, NY, USA, 2009.
18. The United States Environmental Protection Agency (EPA). What Are the Air Quality Standards for PM. Available online: https://www3.epa.gov/region1/airquality/pm-ne-nattain.html (accessed on 10 January 2022).
19. Moran, P. The interpolation of statistical maps. *J. R. Stat. Soc. B* **1948**, *10*, 243–251. [CrossRef]
20. Chi, G.; Zhu, J. *Spatial Regression Models for Social Sciences*; SAGE: New York, NY, USA, 2020.
21. Anselin, L. Local indicators of spatial association—LISA. *Geogr. Anal.* **1995**, *27*, 93–115. [CrossRef]
22. Chowdhuiry, R.I.; Hasan, M.T. Markov regression model for analyzing big data to predict trajectories of repeated categorical outcomes: An application to $PM_{2.5}$ air pollution data. *Environ. Ecol. Stat.* **2022**, *29*, 149–184. [CrossRef]
23. Islam, M.A.; Chowdhury, R.I. *Analysis of Repeated Measures Data*; Springer: Singapore, 2017.
24. Crespo, R.; Manuschevich, D. Using spatial autologistic regression for predicting urban growth. *J. Spat. Sci.* **2023**, *68*, 651–666. [CrossRef]
25. Besag, J. Spatial interaction and statistical analysis of lattice system. *J. R. Stat. Soc. B* **1974**, *36*, 192–236. [CrossRef]
26. Crase, B.; Liedloff, A.C.; Wintle, B.A. A new method for dealing with residual spatial autocorrelation in species distribution models. *Ecography* **2012**, *35*, 879–888. [CrossRef]
27. Sherman, M.; Apanasovich, T.V.; Carroll, R.J. On estimation in binary autologistic spatial models. *Environ. Ecol. Stat.* **2006**, *76*, 167–179. [CrossRef]

28. Wu, H.; Huffer, F.W. Modelling the distribution of plant species using the autologistic regression model. *Environ. Ecol. Stat.* **1997** *4*, 31–48. [CrossRef]
29. Bo, Y.C.; Song, C.; Wang, J.F.; Li, X.W. Using an autologistic regression model to identify spatial risk factors and spatial risk patterns of hand, foot and mouth disease (HFMD) in Mainland China. *BMC Public Health* **2014**, *14*, 358. [CrossRef]
30. Thanapongtharm, W.; Linard, C.; Pamaranon, N.; Kawkalong, S.; Noimoh, T.; Chanachai, K.; Parakgamawongsa, T.; Gilbert, M. Spatial epidemiology of porcine reproductive and respiratory syndrome in Thailand. *BMC Vet. Res.* **2014**, *10*, 174. [CrossRef] [PubMed]
31. Wu, W.; Zhang, L. Comparison of spatial and non-spatial logistic regression models for modeling the occurrence of cloud cover in north-eastern Puerto Rico. *Appl. Geogr.* **2013** *37*, 52–62. [CrossRef]
32. Schultz, C.; Alegría, C.C.; Cornelis, J.; Sahli, H. Comparison of spatial and aspatial logistic regression models for landmine risk mapping. *Appl. Geogr.* **2016**, *66*, 52–63. [CrossRef]
33. Zhang, D.; Ren, N.; Hou, X. Improved logistic regression model based on a spatially weighted technique (ILRBSWT v1.0) and its application to mineral prospectivity mapping. *Geosci. Model Dev.* **2018**, *11*, 2525–2539. [CrossRef]
34. Dormann, C.F. Assessing the validity of autologistic regression. *Ecol. Model.* **2007** *207*, 234–242. [CrossRef]
35. Augustin, N.H.; Augustin, M.A.; Buckland, S.T. An Autologistic Model for the Spatial Distribution of Wildlife. *J. Appl. Ecol.* **1996**, *33*, 339–347. [CrossRef]
36. Gégout-Petit, A.; Guérin-Dubrana, L.; Li, S. A new centered spatio-temporal autologistic regression model with an application to local spread of plant diseases. *Spat. Stat.* **2019**, *31*, 100361. [CrossRef]
37. Wolters, M.A. Better Autologistic Regression. *Front. Appl. Math. Stat.* **2017**, *3*, 100361. [CrossRef]
38. Lindsey, J.K.; Lambert, P. On the appropriateness of marginal models for repeated measurements in clinical trials. *PStat Med.* **1998**, *28*, 447–469. [CrossRef]
39. Lee, Y.; Nelder, J.A. Conditional and Marginal Models: Another View. *Stat. Sci.* **2004**, *19*, 219–238. [CrossRef]
40. Chowdhury, R.I.; Islam, M.A. Regressive models for risk prediction of repeated multinomial outcomes: An illustration using Health and Retirement Study data. *Biom. J.* **2020**, *62*, 898–915. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.