



Published in final edited form as:

Mathematics (Basel). 2025 ; 13(2): . doi:10.3390/math13020180.

Estimating the Relative Risks of Spatial Clusters Using a Predictor-Corrector Method

Majid Bani-Yaghoub¹, Kamel Rekab^{1,*}, Julia Pluta¹, Said Tabharit¹

¹Division of Computing, Analytics and Mathematics, School of Science and Engineering, University of Missouri-Kansas City, Kansas City, MO 64110, USA.

Abstract

Spatial, temporal, and space-time scan statistics can be used for geographical surveillance, identifying temporal and spatial patterns, and detecting outliers. While statistical cluster analysis is a valuable tool for identifying patterns, optimizing resource allocation, and supporting decision-making, accurately predicting future spatial clusters remains a significant challenge.

Given the known relative risks of spatial clusters over the past k time intervals, the main objective of the present study is to predict the relative risks for the subsequent interval, $k + 1$. Building on our prior research, we propose a predictive Markov chain model with an embedded corrector component. This corrector utilizes either multiple linear regression or exponential smoothing method, selecting the one that minimizes the relative distance between observed and predicted values in the k -th interval.

To test the proposed method, we first calculated the relative risks of statistically significant spatial clusters of COVID-19 mortality in the U.S. over seven time intervals from May 2020 to March 2023. Then, for each time interval, we selected the top 25 clusters with the highest relative risks and iteratively predicted the relative risks of clusters from intervals three to seven. The predictive accuracies ranged from moderate to high, indicating the potential applicability of this method for predictive disease analytics and future pandemic preparedness.

Keywords

COVID-19; Mortality; Relative Risk; Cluster Analysis; Exponential Smoothing; Regression

1 Introduction

The emergence of COVID-19 highlighted the critical need for precise spatial and temporal surveillance of infectious diseases, especially in densely populated or vulnerable regions. The spatial scan statistic has proven essential for identifying disease clusters in public health data and helping authorities detect, respond to, and mitigate outbreaks. It was first proposed by Kulldorff [31] and subsequently developed in various studies [33, 26]. This method, using circular and, later, elliptic scanning windows [32], efficiently identifies clusters without prior information on the disease's distribution, providing a robust framework for

*Corresponding Author: rekabk@umkc.edu.

various applications from infectious diseases to cancer epidemiology [8, 5]. Kulldorff's elliptic spatial scan statistics address the limitations of circular windows, particularly in urban environments where population density and movement patterns are uneven [32].

Additional advancements include the weighted normal spatial scan statistic, which accommodates heterogeneous population data, enabling more precise detection in demographically diverse regions [26, 45]. Studies on COVID-19, for instance, have applied these techniques to uncover spatial variations and emerging hot spots within urban settings, providing critical insights into the pandemic's dynamics [2, 3, 4, 6, 34]. Another study, [16] explored how these spatial methods facilitated response coordination across countries, reflecting the adaptability of spatial scan statistics in different geographic contexts and disease typologies. Additionally, the multivariate scan statistic extends traditional models to analyze multiple data types simultaneously, thereby improving surveillance accuracy in multifactorial diseases [33]. The utility of Monte Carlo hypothesis testing, as demonstrated in [14] and further refined for power analyses by [41], is integral to validating spatial clusters, adding another layer of reliability to these models.

These spatial methodologies have also shown efficacy in tracking non-COVID infectious diseases. For example, the space-time models applied airborne diseases like H7N9 influenza [43, 9, 50] demonstrated their utility in understanding disease spread and preparing targeted interventions. Furthermore, Souris et al. [42] illustrated the application of these tools in poultry farms, identifying avian influenza re-emergence risks across human and animal health sectors (see also [12, 21]).

In healthcare-associated infection settings, these models have helped detect and control outbreaks, as shown in studies on methicillin-resistant *Methicillin-resistant Staphylococcus aureus* [27] and carbapenemase-producing *Klebsiella pneumoniae* [1]. By incorporating real-time data, these systems provide timely alerts, allowing healthcare providers to implement interventions more rapidly [36]. The syndromic surveillance adaptations, which emerged from the anthrax preparedness studies of [30] and subsequent research by [25], underscore the proactive role these systems play in modern epidemiological surveillance.

The spatial scan statistic and its derivatives have been instrumental across public health, responding to infectious diseases, environmental exposures, and even syndromic surveillance scenarios. These tools can become even more powerful when applied for prediction, as predictive applications extend the utility of spatial cluster analysis beyond retrospective assessments. The main question is whether predictive methods can be merged with these tools to identify regions at heightened risk of future disease spread, anticipate healthcare resource requirements, and guide targeted interventions in real time, enabling decision-makers to allocate resources more effectively. In this paper, we aim to test the applicability of a predictor-corrector method to the relative risks resulting from the spatial cluster analysis of k time intervals for predicting the relative risks in the $k + 1 - th$ time interval. In particular, assuming that the relative risks of spatial clusters are known for the past k time intervals, we are interested in predicting the relative risks of spatial clusters in the time interval $k + 1$. Based on our previous work [46], we propose a Markov chain model with an embedded corrector. The corrector is either a multiple linear regression or an

exponential smoothing, whichever with shorter relative distance between the observed and predicted values in interval k .

Predicting relative risks in spatial clusters could serve as an early warning system for areas at increased mortality risk, which is especially valuable for addressing resource allocation and intervention planning in critical periods. We use the U.S. COVID-19 mortality data [44] to evaluate this approach and assess whether this method can accurately forecast relative risks in clusters of significant mortality.

The remainder of this paper is organized as follows. Section 2 provides details on the data, the SaTScan method [15], and the results of our spatial cluster analysis. Section 3 introduces the predictor-corrector method, while Section 4 applies this method to the U.S. COVID-19 mortality data over six time intervals, with the goal of predicting relative risks of the seventh time interval. Finally, Section 5 discusses the results and limitations of the present work, as well as potential directions for future research.

2 Retrospective Analysis of Spatial Clusters

2.1 Overview

We employed a retrospective spatial analysis of COVID-19 clusters across the United States by examining how case distributions correlate with population density. Utilizing a Poisson-based spatial scanning model, we identified clusters by employing a cylindrical scanning window to capture areas of heightened risk relative to surrounding regions. The analysis applies likelihood ratio testing to detect significant spatial clusters, with primary and secondary clusters reported based on statistical significance. Monte Carlo simulations were used to validate the results, offering insights into the geographic spread and demographic impact of COVID-19 across different communities in the U.S.

2.2 US COVID-19 Mortality Data

We obtained COVID-19 mortality data for the United States from *The New York Times*, which provides daily records of COVID-19 deaths from February 2020 onward. This data, publicly available at [44].

The dataset reports daily COVID-19 deaths across all U.S. states and territories. In many states, health departments reconcile these records with death certificates, omitting deaths where COVID-19 was not the cause of death. Following this protocol, non-COVID-19 deaths were removed from the dataset when there was a confirmation that COVID-19 was not the cause, specifically in cases of homicide, suicide, car accidents, or drug overdoses. Details of these adjustments can be found in [44].

To examine significant variations in COVID-19 mortality trends, we focused on data spanning from May 24, 2020, to March 12, 2023. We divided this period into seven time intervals to analyze trends within consistent segments of time: May 24, 2020 – September 13, 2020; September 13, 2020 – March 14, 2021; March 14, 2021 – June 13, 2021; June 13, 2021 – October 31, 2021; October 31, 2021 – March 13, 2022; March 13, 2022 – October 16, 2022; and October 16, 2022 – March 12, 2023. Each interval represents a critical phase

in the pandemic, allowing us to study temporal fluctuations in COVID-19-related mortality rates (see Table 1 for a summary). Specifically, these intervals were selected based on the emergence of new *SARS-CoV-2* variants, which played a pivotal role in shaping the trajectory of the pandemic. This approach aimed to capture virological and epidemiological shifts corresponding to significant phases of the outbreak. As shown in Figure 2 (a), each interval, except the third, includes at least one peak in COVID-19 mortality within the United States. The third interval corresponds to the emergence of the Alpha variant, which did not result in a mortality peak during this period but contributed to a significant peak in the subsequent interval. While this method highlights variant-driven dynamics, we acknowledge that alternative approaches, such as structural change detection methods [22], could potentially offer a data-driven determination of breakpoints. Incorporating such methods in future studies could further refine the interval selection process and enhance the robustness of predictive modeling.

2.3 Spatial Statistical Scan

To analyze the geographical distribution of COVID-19 cases adjusted for population demographics, we applied a Poisson model. Our spatial analysis utilized a cylindrical scanning approach, with a moving circular base representing the spatial dimension, and set a maximum cluster size of 25% of the population at risk to prevent excessively large clusters. This approach created numerous overlapping windows of varying sizes, allowing full coverage of the study area. Each window was considered as a potential cluster. Under the null hypothesis H_0 , the COVID-19 risk is uniform inside and outside the scanning window, while the alternative hypothesis H_a suggests a higher risk within the window. Each cylinder expanded up to a preset upper limit on cluster size. To identify clusters, we applied a likelihood ratio test.

The likelihood ratio $\frac{L(C)}{L_0}$ is calculated by

$$\frac{L(C)}{L_0} = \frac{\left(\frac{n_c}{\mu(c)}\right)^{n_c} \left(\frac{N - n_c}{N - \mu(c)}\right)^{N - n_c}}{\left(\frac{N}{\mu(T)}\right)^N}, \quad (1)$$

where $L(C)$ is the maximum likelihood for the cylinder based on the observed cases n_c and expected cases $\mu(c)$ within the window. Here, L_0 is the likelihood under the null hypothesis, N is the total cases observed in the U.S. for each time interval, and $\mu(T)$ represents the total expected cases. A likelihood ratio greater than one indicated that observed cases exceeded expectations. The cluster with the highest likelihood was labeled the primary cluster, while other clusters were reported if they achieved significance at $p < 0.05$. P-values for space-time clusters were determined through Monte Carlo simulations, using 999 randomized iterations of the data to assess statistical significance [15].

2.4 Identified Significant Clusters

Figure 1 illustrates the significant spatial clusters of COVID-19 mortality across the United States. High-risk clusters, characterized by a relative risk greater than one, are represented by red circles, while low-risk clusters, with a relative risk less than one, are indicated by blue circles. Panel (a) shows the significant spatial clusters for the entire study period; however, it may not adequately reflect the temporal changes across the defined time intervals from 1 to 7. Panels (b) through (h) correspond to the spatial data associated with intervals 1 through 7, respectively. Panel (b) shows that significant high-risk clusters were initially concentrated in the southern regions of the U.S. As shown in panel (c), these high-risk clusters extend across much of the country, with the notable exceptions of certain areas in the East and Northwest, which may be attributed to lower vaccination rates. Panel (d) illustrates the impact of extensive vaccination efforts, which have proven to be substantially effective. The emergence of the Delta variant is highlighted in panel (e), demonstrating its influence on the overall mortality trends. The effectiveness of vaccination updates and booster shots is evident in panel (f) when it is compared with panel (e). Panel (g) indicates that the significance of the Omicron variant has become pronounced in the central regions of the U.S. Finally, panel (h) reveals a marked reduction in the prevalence of significant high-risk clusters, underscoring the impact of vaccination and public health interventions on COVID-19 mortality rates.

Figure 2 shows the dynamics of COVID-19 mortality trends and the geographical distribution of risk clusters in the U.S., highlighting key patterns in mortality fluctuations, spatial risk persistence, and transitions between risk classifications. Panel (a) illustrates the time series of COVID-19-related mortality in the U.S. from May 24, 2020, to March 12, 2023, segmented into seven distinct periods. Panel (b) shows the percentage of the U.S. area covered by significant high-risk clusters (i.e., relative risk greater than 1) and low-risk clusters (i.e., relative risk less than 1) for each of the seven intervals. It suggests that both high-risk and low-risk clusters may gradually reach an equilibrium below 20% of the U.S. area, indicating a stable distribution of risk across the country. Panel (c) examines the persistence of high-risk and low-risk clusters from time interval i to interval $i + 1$. It demonstrates that approximately 45% of areas classified as high-risk and about 40% of areas classified as low-risk maintain their status, particularly from time interval 5 onward, suggesting a notable consistency in spatial risk distribution over time. Panel (d) illustrates the percent transition of clusters from high-risk to low-risk and vice versa between consecutive intervals. This transition appears to stabilize around 15%, indicating a moderate but consistent change in the risk classification of certain areas over the specified periods.

Table 2 presents a detailed summary of spatial high-risk and low-risk clusters of COVID-19 mortality across intervals 1 through 7. The second and third columns list the number of high-risk and low-risk clusters identified in each interval. Columns 4 through 7 provide the percentage of the total U.S. area covered by these clusters. Notably, in interval 4, high-risk clusters encompassed more than 50% of the U.S. area.

Table 3 details the transitions between high-risk and low-risk COVID-19 mortality clusters, as well as the persistence of clusters across intervals from i to $i + 1$. Two percentages are

provided: the first represents the overlap of clusters from interval i with those in interval $i + 1$, and the second indicates the overlap of clusters in interval $i + 1$ with interval i . Notably, the values approach steady-state levels over time.

Figure 3 represents the accuracy of identified spatial clusters. Panel (a) shows the observed-to-expected mortality counts plotted against the estimated relative risk for intervals 1–7; panel (b) provides box plots of residuals, indicating low residuals; panel (c) illustrates the mortality data within each identified cluster, giving a localized perspective on mortality distribution; and panel (d) displays the estimated relative risks associated with spatial clusters for each interval, showing the changes in risk levels over time. The observed-to-expected mortality counts and the relative risks of significant spatial clusters are available in the supplementary document Tables 4 and 5, respectively.

3 Modeling Mortality Risk Prediction

3.1 Markov Chain Model

In tackling the challenge of integrating the entire cluster search and prediction process into a manageable model grounded in rigorous statistical analysis, we propose that analyzing trends in transition probabilities across observed clusters over k intervals could enhance the accuracy of our predictions for interval $k + 1$. Given that k is relatively small (between 5 and 10), we employ a predictor-corrector method. The predictor utilizes a Markov chain model [29, 47], while the corrector optimally chooses between multiple linear regression and exponential smoothing methods based on the distance between transition matrices, aiming for an effective prediction of clusters in interval $k + 1$.

Our approach starts with the establishment of initial datasets, labeled T_1 , each representing a specific time interval. As the pandemic progresses, additional datasets T_2, T_3, T_4 , and so on are created, forming a sequence that captures different phases of the pandemic. The collection T_{k+1} denotes the next set of real data gathered at a subsequent interval. Our goal is to predict the parameters of this new collection T_{k+1} using insights from the preceding datasets $T_1, T_2, \dots, T_k$.

Traditionally, predictions for mortality rates have relied solely on the most recent dataset, T_k . However, we recognize that the characteristics of the upcoming dataset T_{k+1} may differ from those represented by T_k . To improve predictive accuracy, we advocate for a broader analysis of the trends in transition probabilities observed across the series of datasets $T_1, T_2, \dots, T_k$.

Understanding the patterns and shifts in transition probabilities over time can provide valuable insights into the evolving dynamics of the pandemic. By examining how these probabilities change, we aim to capture underlying trends that may enhance the accuracy of predicting the parameters for the next dataset, T_{k+1} . This approach contrasts with conventional methods, suggesting that a more nuanced analysis of historical data could lead to a more robust and reliable predictive model for mortality rates in the ongoing fight against COVID-19.

3.2 Estimating the parameters of T_{k+1}

Let T_{k+1}^* represent the prediction for T_{k+1} . Predicting T_1 is impractical because there is no initial data available before T_1 . Similarly, predicting T_2 is challenging due to the presence of only one preceding data set, which does not provide sufficient patterns or trends. However, T_k^* can be predicted for $k > 2$ using the data sets $T_{k-1}, \dots, T_1$. This means that historical data can be maintained as prediction chains throughout the data collection process. More importantly, since predictions are made for each dataset as it becomes available, any inaccurate predictions can be discarded, offering valuable feedback on the quality of the predictive process.

Our proposed method corrects the Markov chain prediction of T_{k+1} utilizing either multiple linear regression or exponential smoothing at each iteration k . The predictor, denoted as T_{k+1}^* , is formulated as follows:

$$T_{k+1}^* = \alpha^* T_k + (1 - \alpha^*) T_k^*, \quad (2)$$

which is a weighted sum between a “good corrector” T_k^* and data T_k . The “good corrector” T_k^* of T_k based on transition matrices from previous datasets $T_1, T_2, \dots, T_{k-1}$, whereas the predictor is given by $\alpha^* T_k$.

Specifically, let $\hat{T}_k$ be a predictor of T_k determined through multiple linear regression, and let $\tilde{T}_k$ be a predictor of T_k determined via exponential smoothing. $\tilde{T}_k$ depends on α^* , an optimal value satisfying $0 < \alpha^* < 1$, and is determined by

$$\sum_{i=3}^k d(\tilde{T}_i(\alpha^*), T_i) = \min_{0 < \alpha < 1} \left\{ \sum_{i=3}^k d(\tilde{T}_i(\alpha), T_i) \right\}, \quad (3)$$

T_k^* is chosen such that

$$d(T_k^*, T_k) = \min \{ d(\hat{T}_k, T_k), d(\tilde{T}_k, T_k) \}. \quad (4)$$

where $d(T_k^*, T_k)$ denotes a distance from transition matrix T_k^* to matrix T_k . We will use a distance measure of the form

$$d(T_k^*, T_k) = \sum_m \sum_n \frac{|(T_k^*(m, n) - T_k(m, n))|}{T_k(m, n) + 1} \quad (5)$$

Here, m represents the matrix row index, while n denotes the matrix column index. This distance metric accounts for relative differences rather than absolute differences. See Ref. [38] for more details.

As a result, we derive the parameters of T_{k+1}^* by utilizing the previous dataset T_k , which is established with real data, and T_k^* , a predictor for T_k that encapsulates historical information from the datasets $T_1, \dots, T_{k-1}$. Throughout this process, we eventually obtain the actual T_{k+1} and compare it to our estimate T_{k+1}^* using the aforementioned distance measure. This method allows us to assess the accuracy of our predictions with each successive data collection. Importantly, the distance measure can reveal whether certain datasets are outliers for various reasons, enabling us to exclude those data points from our analysis.

The ultimate goal is to determine T_{k+1}^* , for which no corresponding T_{k+1} will be available, especially when T_k represents the final data collection before making a prediction. Thus, T_{k+1}^* serves as our prediction for practical use, representing our estimate in the absence of actual observed data for T_{k+1} .

3.3 Predictors of the transition matrix T_{k+1}^*

We employ one of two methods to predict T_{k+1}^* based on the transition matrices $T_1, T_2, \dots, T_k$. The value $\tilde{T}_k$ is calculated using exponential smoothing, while $\hat{T}_k$ is obtained through multiple linear regression. The method that results in the smaller distance measure, as defined earlier, is chosen for the prediction.

3.3.1 Derivation of $\tilde{T}_k$ (exponential smoothing): $\tilde{T}_k$ is determined through the recursion

$$\begin{aligned} \tilde{T}_k &= \alpha T_{k-1} + (1 - \alpha) \tilde{T}_{k-1} \\ \tilde{T}_{k-1} &= \alpha T_{k-2} + (1 - \alpha) \tilde{T}_{k-2} \\ &\vdots \\ \tilde{T}_3 &= \alpha T_2 + (1 - \alpha) \hat{T}_2 \end{aligned} \tag{6}$$

where $\hat{T}_2$ is an estimate of T_2 determined by simple linear regression.

3.3.2 Derivation of $\hat{T}_k$ (multiple linear regression): The general form of $\hat{T}_k$ is

$$\hat{T}_k = a_0 + \sum_{i=0}^{k-1} a_i T_i + \sum_{i < j} \sum_{j=2}^k a_{ij} T_i T_j, \tag{7}$$

where a_i for $i = 1, \dots, k-1$ and a_{ij} for $i < j, j = 1, \dots, k$ are determined through the least square method. In certain cases, the value of the distance is not increased significantly by including

the interaction terms $T_i T_j$. In these cases, a simple multiple linear regression model of the form

$$\hat{T}_k = a_0 + \sum_{i=0}^{k-1} a_i T_i \quad (8)$$

may be used.

4 Predicting Relative Risks of Clusters

In this section, we use the relative risk data obtained from the statistical spatial scan (see Section 2) to test the accuracy of our proposed model. In this process, we aim to use the first six datasets of relative risks to predict the seventh set of relative risks associated with interval seven. Note that we only apply our method to the highest 25 relative risks of each interval. Since the seventh dataset is known, we can evaluate the accuracy of our predictions. Given $T_1, T_2, \dots, T_6$ as the six datasets corresponding to the observations, we seek to determine T_7^* as an appropriate predictor for T_7 . In addition, all coefficients included in the forecasting model were tested for statistical significance, and variables with non-significant coefficients were excluded to ensure the robustness and reliability of the model.

4.1. Exponential Smoothing:

As described in Section 3, the predictor of chain T_6 has the form

$$\begin{aligned} \tilde{T}_6 &= \alpha T_5 + (1 - \alpha) \tilde{T}_5 \\ &\cdot \quad \cdot \quad \cdot \\ &\cdot \quad \cdot \quad \cdot \\ &\cdot \quad \cdot \quad \cdot \\ \tilde{T}_3 &= \alpha T_2 + (1 - \alpha) \hat{T}_2 \\ \hat{T}_2 &= 1.568 + 0.131 T_1 \end{aligned}$$

where $\hat{T}_2$ is a predictor of T_2 , given T_1 , determined through simple linear regression. $\tilde{T}_6(\alpha^*)$ is chosen so that

$$\sum_{i=3}^6 d(\tilde{T}_i(\alpha^*), T_i) = \min_{0 < \alpha < 1} \sum_{i=3}^6 \{d(\tilde{T}_i(\alpha), T_i)\},$$

where the distance d is defined as in Section 3. represents the values of $D(\alpha) = \sum_{i=3}^6 d(\tilde{T}_i(\alpha), T_i)$. The minimum occurs for $\alpha^* = 0.02$ giving

$$\tilde{T}_6 = 0.02 T_5 + 0.98 \hat{T}_5$$

Also, note that

$$d(\tilde{T}_6, T_6) = \sum_{m=1}^6 \sum_{n=1}^6 \frac{|\tilde{T}_6(m, n) - T_6(m, n)|}{T_k(m, n) + 1} = 11.756$$

4.2 Multiple Linear Regression:

Multiple linear regression relates transition matrix T_6 to $T_1, \dots, T_6$ in one linear functional form. Results in least squares coefficients are presented in Table 2.

The prediction equation is

$$\hat{T}_6 = 0.5T_1 + 6.357T_2 - 0.561T_3 + 12.287T_4 - 8.645T_5$$

$$d(\hat{T}_6, T_6) = \sum_{m=1}^6 \sum_{n=1}^6 \frac{|\hat{T}_6(m, n) - T_6(m, n)|}{T_k(m, n) + 1} = 5.887$$

which is smaller than $d(\tilde{T}_6, T_6)$. The better predictor, therefore, of T_6 is $\hat{T}_6$. Consequently, the predictor of T_7 is

$$T_7^* = 0.02T_6 + 0.98\hat{T}_6 \quad (9)$$

Figure 4 shows the estimation of the optimal parameter α , and the prediction accuracy of relative risks for interval 7 across different models. Panel (a) shows the determination of the optimal α by minimizing the distance between estimated and observed relative risk values. Panel (b) compares observed versus predicted values (on a log scale) for the proposed Markov chain model and the Exponential Smoothing. The predictions of the relative risks look promising but have room for improvement. The proposed model demonstrates superior predictive performance compared to the Exponential Smoothing method, as evidenced by a lower Mean Squared Error (164.8 vs. 258.7).

5 Discussion

This study represents an initial step toward enabling statistical spatial cluster analysis to predict future spatial clusters effectively. By applying our developed method to U.S. COVID-19 mortality data across seven time intervals, we demonstrate that predictive accuracy ranges from moderate to high (Figure 4), indicating its potential utility in disease surveillance, resource allocation, and pandemic preparedness. The length of time intervals can be adapted based on the characteristics of specific infectious diseases and data availability. Despite these promising findings, several limitations warrant consideration.

First, it is important to recognize that predicting relative risks for subsequent intervals does not ensure that spatial clusters will consistently align with those of prior intervals. As shown in Table 3, the overlap between high-risk regions from one interval to the next (i.e., interval $i+1$ and interval i) is often limited, varying between 17% and 51%. Nevertheless, these

relative risk predictions can provide health officials and policymakers with an anticipatory perspective on likely future conditions, offering a gauge of infection severity that may aid in decision-making.

Second, the accuracy of the correction phase in the Markov chain model could be enhanced by incorporating additional techniques, such as autoregressive or moving average models, or by accounting for seasonality when relevant, as with influenza. Although the current model's relative risk predictions show acceptable accuracy, future research should focus on refining these predictions to improve reliability.

Third, and perhaps most critically, the effectiveness of relative risk predictions would be substantially enhanced by integrating predictions of spatial cluster radii. Our attempts to incorporate radii into the same predictive matrices as relative risks did not yield high accuracy. We propose that future models should separate radius and relative risk predictions into two interconnected submodels to improve prediction precision and practical applicability.

Our Future research will aim to advance the proposed methodology and enhance its applicability by addressing a few key areas as follows. Overcoming inconsistent spatial cluster alignment remains a significant challenge, as variations between consecutive intervals can limit the reliability of predictions. Developing algorithms that dynamically adjust the time intervals to overcome inconsistencies, such as machine learning-based alignment models [28], could significantly improve the robustness of our methodology. Additionally, enhancing the accuracy of the correction phase through hybrid models that combine Markov chains with differential equations [10], neural networks [48] or ensemble techniques [20] could improve the predictions. Furthermore, integrating spatial cluster radii predictions with relative risks in a cohesive framework will require innovative approaches, such as leveraging spatiotemporal modeling [49] and advanced simulation techniques [17], to achieve precise and reliable outputs. These enhancements could broaden the scope of applications for disease surveillance, calculation of the basic reproduction number R_0 for each spatial cluster [11], and health resource management, ultimately contributing to more effective public health interventions.

In conclusion, this study highlights the potential of modified Markov chain models in predicting the relative risks of spatial clusters. This approach could be further developed to predict not only the relative risks but also the location and radius of each spatial cluster, enhancing its value for public health planning and intervention.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

This study was partially supported by the Center for Disease Control and Prevention under grant number 5U01CK000671-02.

References

- [1]. Abboud CS et al. A space-time model for carbapenemase-producing klebsiella pneumoniae in hospitals. *Epidemiology and Infection*, 2015.
- [2]. AlQadi H and Bani-Yaghoub M. Incorporating global dynamics to improve the accuracy of disease models: Example of a covid-19 sir model. *Plos one*, 17(4):e0265815, 2022. [PubMed: 35395018]
- [3]. AlQadi H, Bani-Yaghoub M, Balakumar S, Wu S, and Francisco A. Assessment of retrospective covid-19 spatial clusters with respect to demographic factors: case study of kansas city, missouri, united states. *International Journal of Environmental Research and Public Health*, 18(21):11496, 2021. [PubMed: 34770012]
- [4]. AlQadi H, Bani-Yaghoub M, Wu S, Balakumar S, and Francisco A. Prospective spatial-temporal clusters of covid-19 in local communities: case study of kansas city, missouri, united states. *Epidemiology & Infection*, 151:e178, 2023.
- [5]. Amin S et al. Applications of spatial scan statistics in disease cluster detection. *International Journal of Health Geographics*, 2014.
- [6]. Andrade LA et al. Covid-19 mortality in an area of northeast brazil. *Epidemiology & Infection*, 148, 2020.
- [7]. Anwar A, Malik M, Raees V, and Anwar A. Role of mass media and public health communications in the covid-19 pandemic. *Cureus*, 12(9), 2020.
- [8]. Rikke Baastrup N et al. Use of spatial scan statistics for cancer epidemiology. *Environmental Research*, 2013.
- [9]. Badaloni C et al. Space-time clustering of covid-19 in europe. *Geospatial Health*, 2020.
- [10]. Bani-Yaghoub Majid, Elhomani Abdellatif, and Catley Delwyn. Effectiveness of motivational interviewing, health education and brief advice in a population of smokers who are not ready to quit. *BMC Medical Research Methodology*, 18:1–10, 2018. [PubMed: 29301497]
- [11]. Bani-Yaghoub Majid, Gautam Raju, Shuai Zhisheng, Van Den Driessche P, and Ivanek Renata. Reproduction numbers for infections with free-living pathogens growing in the environment. *Journal of biological dynamics*, 6(2):923–940, 2012. [PubMed: 22881277]
- [12]. Baygents G and Bani-Yaghoub M. Cluster analysis of hemorrhagic disease in missouri's white-tailed deer population: 1980–2013. *BMC Ecology*, 18:1–8, 2018. [PubMed: 29347979]
- [13]. Benjamin-Chung J and Reingold A. Measuring the success of the us covid-19 vaccine campaign—it's time to invest in and strengthen immunization information systems. *American Journal of Public Health*, 111(6):1078–1080, 2021. [PubMed: 33600253]
- [14]. Besag J and Clifford P. Sequential monte carlo p-values. *Biometrika*, 78:301–330, 1991.
- [15]. Block R. Software review: scanning for clusters in space and time: a tutorial review of satscan. *Social Science Computer Review*, 25(2):272–278, 2007.
- [16]. Bonnet E. et al. The covid-19 pandemic in francophone west africa. *Research Square*, 2020.
- [17]. Caprarrelli G and Fletcher S. A brief review of spatial analysis concepts and tools used for mapping, containment and risk modelling of infectious diseases and other illnesses. *Parasitology*, 141(5):581–601, 2014. [PubMed: 24476672]
- [18]. Christie A, Mbaeyi SA, and Walensky RP. Cdc interim recommendations for fully vaccinated people: an important first step. *JAMA*, 325(15):1501–1502, 2021. [PubMed: 33688914]
- [19]. CDC Team VBCI Covid, Birhane Martha, Bressler Samantha, Chang Grace, Clark Teresa, ..., and Trujillo Alfonso. Covid-19 vaccine breakthrough infections reported to cdc—united states, january 1–april 30, 2021. *Morbidity and Mortality Weekly Report*, 70(21):792, 2021. [PubMed: 34043615]
- [20]. Dietterich Thomas G.. Ensemble methods in machine learning. *Handbook of Learning and Approximation*, 1:1–16, 2000.
- [21]. Gautam R, Srinath I, Clavijo A, Szonyi B, Bani-Yaghoub M, Park S, and Ivanek R. Identifying areas of high risk of human exposure to coccidioidomycosis in texas using serology data from dogs. *Zoonoses and Public Health*, 60(2):174–181, 2013. [PubMed: 22856539]
- [22]. Ghaderpour Ebrahim, Pagiatakis Spiros D, and Hassan Quazi K. A survey on change detection and time series analysis with applications. *Applied Sciences*, 11(13):6141, 2021.

- [23]. Gostin LO, Cohen IG, and Koplan JP. Universal masking in the united states: the role of mandates, health education, and the cdc. *JAMA*, 324(9):837–838, 2020. [PubMed: 32790823]
- [24]. Gostin LO and Wiley LF. Governmental public health powers during the covid-19 pandemic: stay-at-home orders, business closures, and travel restrictions. *JAMA*, 323(21):2137–2138, 2020. [PubMed: 32239184]
- [25]. Horst MA and Coco AS. Spread of illnesses in a community using gis. *Journal of the American Board of Family Medicine*, 23:32–41, 2010. [PubMed: 20051540]
- [26]. Huang L, Huang L, Tiwari R, Zuo J, Kulldorff M, and Feuer E. Weighted normal spatial scan statistic for heterogeneous population data. *Journal of the American Statistical Association*, 104:886–898, 2009.
- [27]. Huang SS et al. Automated outbreak detection in hospitals. *PLoS Medicine*, 7, 2010.
- [28]. Ji D, Dunn E, and Frahm J. Spatio-temporally consistent correspondence for dense dynamic scene modeling. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VI 14*, pages 3–18. Springer International Publishing, 2016.
- [29]. Kemeny JG and Snell LJ. *Finite Markov Chains*. Springer, New York, 1976.
- [30]. Kleinman K et al. Space-time scan statistic in syndromic surveillance. *Epidemiology and Infection*, 133:409–419, 2005. [PubMed: 15962547]
- [31]. Kulldorff M. A spatial scan statistic. *Communications in Statistics - Theory and Methods*, 26(6):1481–1496, 1997.
- [32]. Kulldorff M, Huang L, Pickle L, and Duczmal L. An elliptic spatial scan statistic. *Statistics in Medicine*, 25:3929–3943, 2006. [PubMed: 16435334]
- [33]. Kulldorff M, Mostashari F, Duczmal L, Yih K, Kleinman K, and Platt R. Multivariate spatial scan statistics for disease surveillance. *Statistics in Medicine*, 26:1824–1833, 2007. [PubMed: 17216592]
- [34]. Leveau CM. Variations of covid-19 mortality in buenos aires. *Scielo Preprints*, 2020.
- [35]. Murthy N. Advisory committee on immunization practices recommended immunization schedule for adults aged 19 years or older—united states, 2022. *MMWR. Morbidity and Mortality Weekly Report*, 71, 2022.
- [36]. Natale A et al. Detection of antimicrobial resistance clusters. *Eurosurveillance*, 22(30484), 2017.
- [37]. Raamkumar AS, Tan SG, and Wee HL. Measuring the outreach efforts of public health authorities and the public response on facebook during the covid-19 pandemic in early 2020: cross-country comparison. *Journal of Medical Internet Research*, 22(5):e19334, 2020. [PubMed: 32401219]
- [38]. Rohatgi VK. *An Introduction to Probability Theory and Mathematical Statistics*. Wiley, New York, 1976.
- [39]. Sen-Crowe B, McKenney M, and Elkbuli A. Social distancing during the covid-19 pandemic: Staying home save lives. *The American Journal of Emergency Medicine*, 38(7):1519, 2020. [PubMed: 32305155]
- [40]. Shekhar R, Garg I, Pal S, Kottewar S, and Sheikh AB. Covid-19 vaccine booster: to boost or not to boost. *Infectious Disease Reports*, 13(4):924–929, 2021. [PubMed: 34842753]
- [41]. Silva I, Assunção RM, and Costa M. Power of the sequential monte carlo test. *Sequential Analysis*, 28:163–174, 2009.
- [42]. Souris M et al. Poultry farm vulnerability and avian influenza risk. *International Journal of Environmental Research and Public Health*, 11:934–951, 2014. [PubMed: 24413705]
- [43]. Tchole AI et al. Epidemic and control of covid-19 in niger. *Journal of Global Health*, 10, 2020.
- [44]. The New York Times. Coronavirus (covid-19) data in the united states, 2021. Retrieved June 21, 2023, from <https://github.com/nytimes/covid-19-data>.
- [45]. Tran Thao, Bani-Yaghoub Majid, and DeLisle James R. Non-emergency responses in the 311 system during the early stage of the covid-19 pandemic: a case study of kansas city. *Disaster Prevention and Resilience*, 2023.
- [46]. Whittaker JA, Rekab K, and Thomason MG. A markov chain model for predicting the reliability of multi-build software. *Information and Software Technology*, 42(12):889–894, 2000.

- [47]. Whittaker JA and Thomason MG. Markov chain model for statistical software testing. *IEEE Transactions on Software Engineering*, 20(10):812–824, 1994.
- [48]. Xiang R, Chen L, and Han X. Hybrid markov chain and neural network models for spatial prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 33(8):3623–3635, 2022.
- [49]. Yang H, Zou Y, Wang Z, and Wu B. A hybrid method for short-term freeway travel time prediction based on wavelet neural network and markov chain. *Canadian Journal of Civil Engineering*, 45(2):77–86, 2018.
- [50]. Zhang Y et al. Avian influenza a (h7n9) cluster analysis. *International Journal of Environmental Research and Public Health*, 12:816–828, 2015. [PubMed: 25599373]

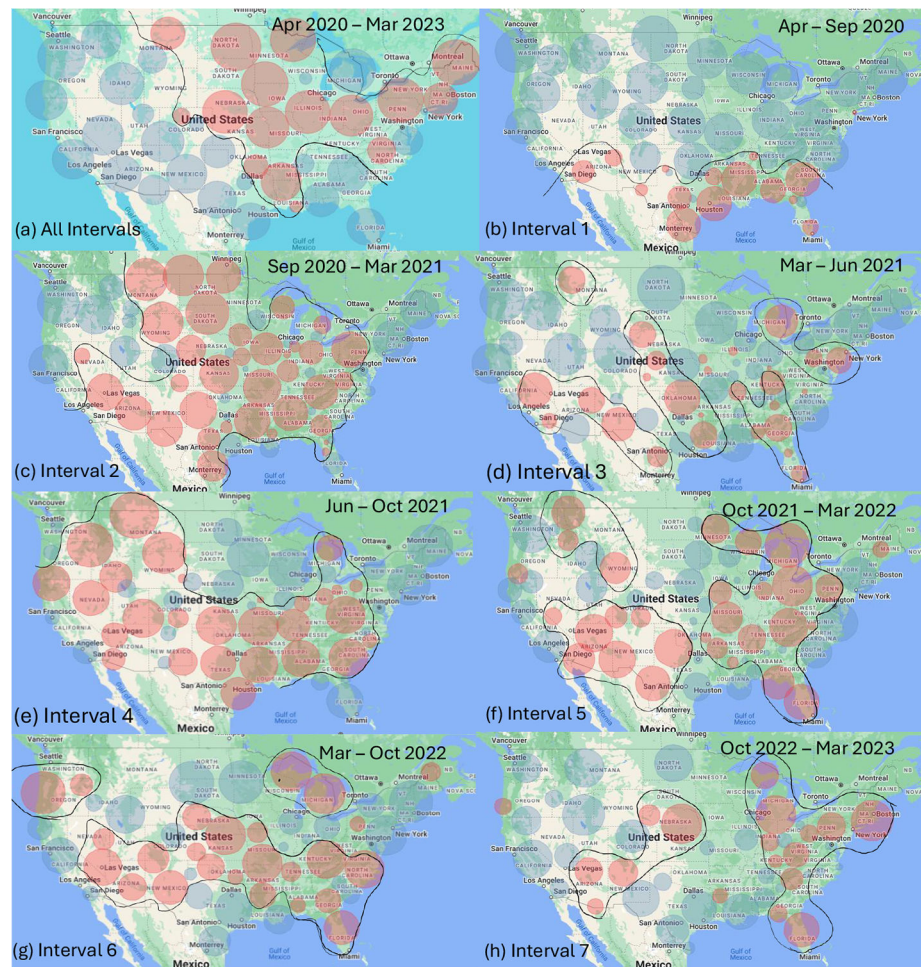


Figure 1: Significant spatial clusters of U.S. COVID-19 mortality. High-risk clusters (relative risk > 1) indicated by red circles, and low-risk clusters (relative risk < 1) shown in blue. (a) the overall significant spatial clusters; (b) the distribution of high-risk clusters was predominantly concentrated in the southern U.S. (c)-(h) notable temporal changes observed across the intervals representing the effects of vaccination efforts (d); the rise of the Delta variant (e), and the increasing significance of the Omicron variant (g), a substantial reduction of high-risk clusters by the end of the study period in panel (h).

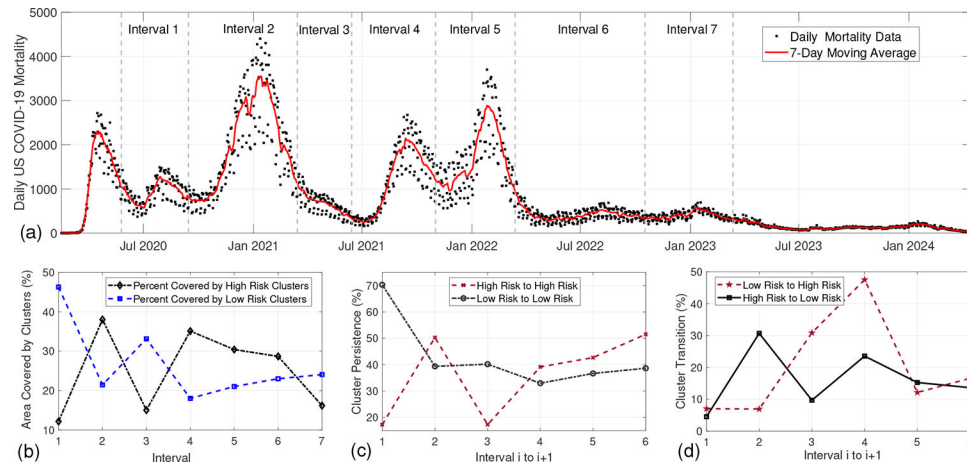
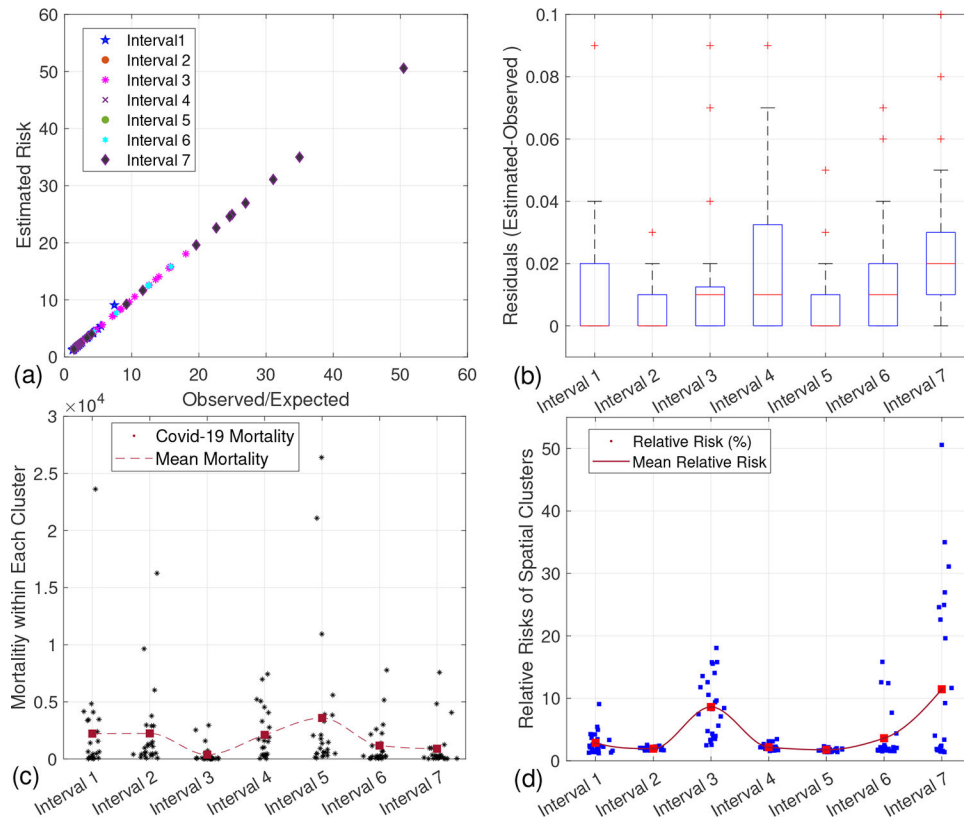
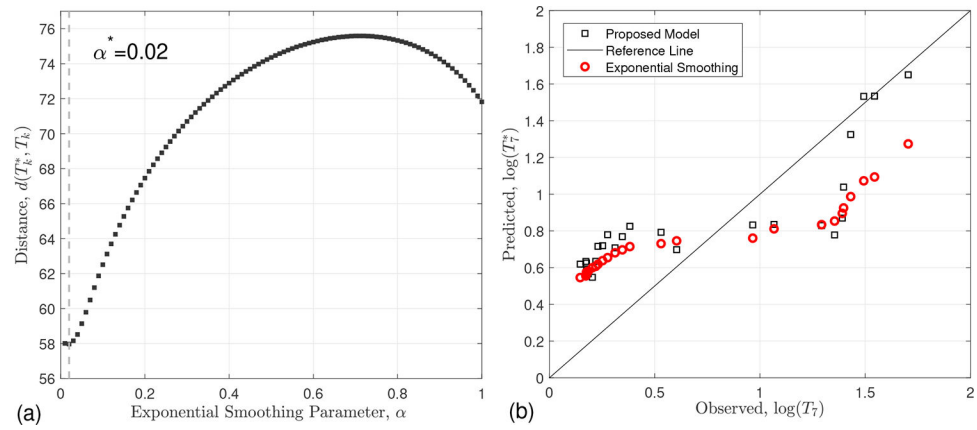


Figure 2:

Time series of mortality data and spatial clusters (a) Time series of U.S. COVID-19 mortality data across seven defined time intervals: (1) May 24, 2020 – September 13, 2020; (2) September 13, 2020 – March 14, 2021; (3) March 14, 2021 – June 13, 2021; (4) June 13, 2021 – October 31, 2021; (5) October 31, 2021 – March 13, 2022; (6) March 13, 2022 – October 16, 2022; (7) October 16, 2022 – March 12, 2023. The figure illustrates trends in mortality over these periods, highlighting significant fluctuations in COVID-19-related deaths; (b) the percentage of U.S. area covered by significant high-risk (i.e. relative risk greater than 1) and low-risk (relative risk less than 1) clusters for each interval of 1 through 7; (c) persistence of high-risk and low-risk clusters from the time interval i to interval $i+1$; (d) percent transition of high-risk to low-risk clusters and vice versa from the time interval i to interval $i + 1$.

**Figure 3:**

Accuracy of identified spatial clusters. (a) observed/expected mortality versus estimated relative risk for intervals 1–7; (b) box plots of residuals, (c) plot of mortality data within each cluster, (d) plot of estimated relative risks of spatial clusters associated with each interval.

**Figure 4:**

Optimal value and the accuracy of predictions for interval 7. (a) Estimating the optimal value of α based on the distance between the estimated and observed relative risk values. (b) log scale observed versus predicted by the proposed Markov chain model for interval 7. The proposed model predictions are superior to those of Exponential Smoothing (Mean Squared Error 164.8 versus 258.7)

Table 1:**US COVID-19 Control and Preventive Measures by Time Interval**

Interval	Variants	Control and Preventive Measures	Ref.
1. 5/24/2020 – 9/13/2020	Original strain (Wuhan)	Social distancing measures, use of face masks, limitations on gatherings, travel restrictions	[39, 24]
2. 9/13/2020 – 3/14/2021	D614G variant	Expanded mask mandates, introduction of COVID-19 vaccines (Pfizer, Moderna), enhanced testing and contact tracing	[23, 13]
3. 3/14/2021 – 6/13/202	Alpha variant (B.1.1.7)	Vaccination campaigns ramped up, continued mask-wearing in high-transmission areas	[19, 18]
4. 6/13/2021 – 10/31/2021	Alpha variant predominant	CDC recommendations for vaccinated individuals, return to in-person learning in colleges and schools, ongoing vaccination efforts	[18, 35]
5. 10/31/2021 – 3/13/2022	Delta variant (B.1.617.2)	Booster shots recommended, reinstated mask mandates in certain areas, increased indoor venue capacity restrictions	[35, 40]
6. 3/13/2022 – 10/16/2022	Omicron variant (B.1.1.529)	Vaccination updates including boosters, recommendations for masks in crowded indoor settings, continued public health surveillance	[39, 35]
7. 10/16/2022 – 3/12/2023	Omicron subvariants (BA.1, BA.2)	Focus on public health awareness campaigns, emphasis on personal responsibility regarding health measures	[7, 37]

Table 2:

Summary of spatial high-risk and low-risk COVID-19 mortality clusters across intervals 1–7, showing the number of clusters and the percentage of U.S. area covered by each.

Interval	High Risk	Low Risk	Area of High Risk	Area of Low Risk	Overlap Area	Total Area
1	25	38	1,170,531.55 (12.80%)	4,457,499.54 (48.73%)	4,830.30 (0.05%)	5,623,200.79 (61.47%)
2	38	34	3,661,450.98 (40.03%)	2,065,006.01 (22.57%)	7,781.26 (0.09%)	5,718,675.73 (62.52%)
3	40	39	1,447,800.92 (15.83%)	3,189,789.31 (34.87%)	5,915.07 (0.06%)	4,631,675.16 (50.63%)
4	30	25	4,667,511.42 (51.03%)	1,733,777.69 (18.95%)	24,583.89 (0.27%)	6,376,705.22 (69.71%)
5	29	36	2,928,563.46 (32.02%)	2,024,862.64 (22.14%)	14,785.95 (0.16%)	4,938,640.15 (53.71%)
6	29	25	2,760,890.45 (30.18%)	2,214,642.66 (24.21%)	3,259.16 (0.04%)	4,972,273.95 (54.36%)
7	28	30	1,556,809.44 (17.02%)	2,316,998.49 (25.33%)	5,328.31 (0.06%)	3,868,479.62 (42.29%)

Table 3:
Transition and persistence rates of high-risk and low-risk COVID-19 mortality clusters across consecutive intervals.

Intervals	High Risk Overlap	Low Risk Overlap	High-Low Transition	Low-High Transition
1 → 2	630,606.00 53.87%, 17.22%	1,451,535.93 32.56%, 70.29%	93,533.80 7.99%, 4.53%	11,311,032.80 29.41%, 35.81%
2 → 3	727,798.55 19.88%, 50.27%	1,255,175.77 60.78%, 39.35%	979,045.90 26.74%, 30.69%	100,420.49 4.86%, 6.94%
3 → 4	583,839.38 40.33%, 17.27%	696,469.49 21.83%, 40.17%	168,040.31 11.61%, 9.69%	1,042,772.11 32.69%, 30.85%
4 → 5	1,146,392.24 33.92%, 39.15%	667,377.54 38.49%, 32.96%	476,609.88 14.10%, 23.54%	1,391,524.37 21.28%, 12.60%
5 → 6	1,178,061.49 40.23%, 42.67%	811,950.67 40.09%, 36.66%	337,960.92 11.54%, 15.26%	335,357.67 16.56%, 12.15%
6 → 7	802,151.60 29.05%, 51.53%	895,447.97 40.43%, 38.65%	315,871.16 11.44%, 13.63%	257,680.39 11.64%, 16.55%