

North American Actuarial Journal

ISSN: 1092-0277 (Print) 2325-0453 (Online) Journal homepage: www.tandfonline.com/journals/uaaj20

Spatial Cyber Loss Clusters at County Level and Socioeconomic Determinants of Cyber Risks

Jaehun Cho, Martin Eling & Kwangmin Jung

To cite this article: Jaehun Cho, Martin Eling & Kwangmin Jung (2025) Spatial Cyber Loss Clusters at County Level and Socioeconomic Determinants of Cyber Risks, North American Actuarial Journal, 29:2, 345-389, DOI: 10.1080/10920277.2024.2408263

To link to this article: https://doi.org/10.1080/10920277.2024.2408263

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC

Published online: 22 Oct 2024.

Submit your article to this journal

Article views: 1609

View related articles

View Crossmark data

Citing articles: 2 View citing articles

Full Terms & Conditions of access and use can be found at https://www.tandfonline.com/action/journalInformation?journalCode=uaaj20

North American Actuarial Journal, 29(2), 345–389, 2025
© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC
ISSN: 1092-0277 print / 2325-0453 online
DOI: 10.1080/10920277.2024.2408263

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Spatial Cyber Loss Clusters at County Level and Socioeconomic Determinants of Cyber Risks

Jaehun Cho,[1] Martin Eling,[2] and Kwangmin Jung[1]
[1]*Department of Industrial and Management Engineering, POSTECH, Pohang, South Korea*
[2]*Institute of Insurance Economics, University of St. Gallen, St. Gallen, Switzerland*

This study investigates whether cyber loss events occurring in the United States are spatially correlated and if so, which socioeconomic factors are associated with the spatial correlation. We analyze 3132 counties of the 50 U.S. states from 2005 to 2020 using the largest existing dataset of cyber risks and socioeconomic data. While previous literature found no or little spatial correlation at the state level, we are the first to document that such correlation exists at the county level; positive Moran's *I* indicates that more exposed (i.e., a relatively large number of cyber events and losses) and less exposed counties are clustered. Spatial regressions show positive direct and negative indirect effects of county-level population and average income on loss frequency and severity. Large and wealthy counties thus tend to be more exposed to cyber risk events, but their geographically neighboring counties are less affected. We further investigate relatively exposed regions (California and the Northeast Coast) and three risk types (malicious, unintended, and privacy risks) and show consistent spatial effects for the key variables of population size and average income. Our findings can aid risk managers, cyber insurers, and policymakers to geographically differentiate cyber risk, recognize relatively more exposed regions, and develop more effective risk management strategies.

JEL classification: C31; C38; G22

1. INTRODUCTION

In recent years, academic papers and industry studies have documented an increase in the frequency and severity of cyber risk events (Edwards et al. 2016; Jung 2021; Ponemon Institute 2022). With respect to severity, billions of data records have been breached in large-scale attacks, such as on Yahoo in 2013 and 2016, on Aadhaar in 2018, and on First American in 2019, all of which caused the loss of sensitive personal information and reputational damage to firms (Vojinovic 2022). With respect to frequency, the Ponemon Institute (2022) considers 550 organizations across 17 countries and highlights that 83% of the studied organizations had experienced a data breach event at least once in the past.

Cyberattacks such as malware, denial of service, or phishing feature borderless occurrence and uncertainty in root causes (Carr 2012). High interdependence and uncertainty, as well as its dynamic nature, are salient characteristics distinguishing cyber risk from traditional risks. Economic agents exposed might be concerned about negative externalities from interdependent parties in neighboring regions. Cyber risk literature has documented that such contagion effects may lead to substantial losses for individuals, industries, and societies (see, e.g., Aldasoro et al. 2022; Eisenbach et al. 2022; Eling and Jung 2022).

However, very little academic literature has discussed the geographical landscape of cyber risks showing potential geographical heterogeneity in cyber loss exposures, where some regions tend to have high exposures to cyber risks, while others do not (Khey and Sainato 2013; Ma, Chu, and Jin 2022). Many practical analytics of cyber risks, though, have identified firmographics of cyber risks and geographical location as critical risk factors (Namestnikov 2012; Sen and Borle 2015; Guy Carpenter 2022). Industry studies report that victim organizations in the United States are spatially clustered in industrial

Address correspondence to Kwangmin Jung, Cheongam-ro 77, Nam-gu, Pohang, Gyeongbuk, South Korea. E-mail: kwjung@postech.ac.kr

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

centers, for example, in the Northeast or Silicon Valley (Windrem 2015). Also, as a major financial hub, Wall Street might be a target due to its technological interconnectivity (Munro 2016).[1]

Some clues to understanding geographical heterogeneity can be found in the criminological context. The criminological literature, especially the widely discussed social disorganization theory, has documented that the physical and social environment plays a role in understanding (cyber)crimes.[2] For example, Kigerl (2012), Sen and Borle (2015), and Park et al. (2019) examine what socioeconomic factors are more likely to drive cybercrime and document that factors such as income, education, and inequality affect cybercrime frequency. Other determinants of exposures might be economic development, Internet penetration, and industrial structure. Such studies in the criminological context help with understanding under what circumstances of a certain region cyber risk events are more likely to occur. However, these studies do not consider the spatial features of such impacts, which may help better understand the socioeconomic determinants affecting neighboring regions and potential regional clusters in frequency and severity of events.

This study investigates whether spatial clusters of cyber risk events can be empirically observed (i.e., regional clusters relatively more exposed to cyber risks) and if so, what the spatial impact of selected socioeconomic factors on cyber risk events looks like. We use the largest existing dataset of cyber risks events provided by Advisen and socioeconomic information of the United States at both the state and county levels.[3] Considering the global and local Moran's I statistics as estimators of spatial autocorrelation, we find no evidence on geographically spatial dependency in both cyber loss frequency and severity at the state level (in line with existing literature), but we do identify the presence of county-level dependency for both (which is, to our best knowledge, a new result). Our findings provide an argument for the government to focus its support on counties that are more exposed to cyber risk events.[4]

Based on the spatial autocorrelation analysis, we implement spatial econometric approaches to investigate the spatial impacts of socioeconomic factors (i.e., average income, population as a proxy for Internet usage, industry identifiers) on cyber loss frequency and severity. We compare three spatial econometric approaches (spatial lag model, spatial error model, and spatial autoregressive combined model) to capture the spatial effects. We find that the population and average income of a county have significant and positive direct, but negative indirect, effects[5] in explaining cyber loss frequency and severity. This implies that large and wealthy counties tend to be more exposed to cyber risk events, but their neighboring counties are less affected by such events.

This study contributes to literature in the following dimensions. To our knowledge, it is the first study to explore geographical cyber loss patterns between counties of the United States and spatial impacts of socioeconomic factors on cyber loss frequency and severity.[6] In addition, we explore a dataset combining cyber risk data with geographical information and

[1]Given that for some types of events attackers and victims are not necessarily in the same place, the geographical dimension for cyber risk might seem not as intuitive as for other types of crimes. However, there are various aspects that indicate a geographical dimension for cyber risks. The Internet infrastructure tends to feature physical spatial network systems (Tranos 2013; Schmidtke 2018). Internet use delays thus generally rely on physical distances between user nodes measured by roundtrip time (Schmidtke 2018). In addition, telecommunications firms tend to construct Internet infrastructure in agglomeration economies or industrial hubs, given considerable construction costs (Malecki 2002; Priemus 2007). The risk exposure as well as the number of potential attackers might thus be clustered. We also note that for many types of events such as a stolen laptop or unintentional disclosure of information, actor and victim are in the same place.

[2]The social disorganization theory claims that a person's residential location is a more significant predictor of criminal activity than the person's characteristics. The routine activity theory in criminology supports this idea by demonstrating that crimes typically occur in an environment where criminals can be motivated to commit crimes, for example, a set of targets without proper protection or a set of targets more likely to be exposed to attackers (Felson 2008; Reyns 2017).

[3]In Appendix C, we implement a robustness test of our empirical results with an alternative public dataset of data breach risks provided by the Privacy Rights Clearinghouse (PRC). The literature on data breach risk analysis has widely used the PRC dataset that provides loss information from 2005 to 2020 (see, e.g., Edwards, Hofmeyr, and Forrest 2016; Eling and Loperfido 2017; Eling and Jung 2018).

[4]In this context, we examine in the spatial panel analysis the state of California and the Northeast Coast area (i.e., part of New York State, New Jersey, Connecticut, and Massachusetts) that are relatively more exposed to cyber risks. We further investigate clusters of three risk types (malicious, unintended, and privacy risks), where we observe significant spatial impacts.

[5]A direct spatial effect indicates the spatial relationship between the explanatory variables and the dependent variable within a region, whereas an indirect spatial effect addresses the relationship across neighboring regions.

[6]Existing studies have documented how network interconnection (i.e., network topology) between cyber exposures can be modeled (see, e.g., Hua and Xu 2021) and whether state-level loss clusters exist by using a clustering approach (see, e.g., Ma et al. 2022). Our study investigates geographical loss patterns at the county level and the spatial impacts of socioeconomic factors on such patterns, which complements and adds to the literature by offering evidence on spatial drivers to better understand microlevel patterns of cyber risk.

socioeconomic data, a combination that to our knowledge has not been used in the literature.[7] We also study spatial features of cyber risks at the county level, which can help one find an answer to what regional and local approaches can be effective to mitigate such risks.[8] Our contributions can be useful for risk managers, cyber insurers, and policymakers to (1) improve the cyber risk analysis and underwriting by differentiating risks geographically and (2) recognize relatively more exposed regions to develop more effective risk management policies.

The rest of the article is structured as follows. In Section 2, we review the literature on cyber risks, followed by the theoretical background on the spatial methods in Section 3. Section 4 describes the dataset used in the study and results of spatial dependency and econometric models are addressed in Section 5. We conclude in Section 6.

2. LITERATURE REVIEW AND DISCUSSION OF POTENTIAL SOCIOECONOMIC FACTORS

There are two research streams with relevance to this study. First, several studies in risk management and insurance have examined statistical features of cyber risks and related loss events, without consideration of the spatial dimension, focusing on firm-specific drivers of cyber risks. Second, a few papers conduct spatial analysis of cyber risk events in information security and more recently also in the insurance literature. These studies have documented heavy-tailedness of cyber losses, the presence of temporal dependence in extreme cyber losses, the presence of spatial dependence, and significant impacts of some socioeconomic factors (e.g., income, population, education) on cyber risk events, which we discuss in this section.

2.1. General Cyber Loss Analysis and Firm-Specific Drivers of Cyber Risk

Edwards et al. (2016) examine data breach frequency and severity and find that the negative binomial and lognormal distributions are the best fit for loss frequency and severity, respectively. Eling and Loperfido (2017) support this finding by showing identical distribution fitting results. They also determine clusters between different risk types and industries using multidimensional scaling (MDS) and with this emphasize heterogeneity between risk types and industries. One example is that retailers face very different cyber risks compared with other industries (e.g., the financial industry and the public sector).

In addition to general features of cyber loss frequency and severity, several studies have investigated extreme cyber loss events. Using a truncated Pareto distribution for extreme cyber risks, Wheatley, Maillart, and Sornette (2016) present that the cumulative sum of data breach losses is expected to double within the next five years, from two to four billion breached records in this period. Eling and Wirfs (2019) claim that cyber risk data should be analyzed with extreme value theory (EVT)-based models and find that the peaks-over-threshold (POT) method with a lognormal for the body of the severity distribution and generalized Pareto for the tail of this distribution can appropriately describe the statistical features of cyber losses. Jung (2021) studies the temporal dependency of extreme data breach losses (maxima for a certain time frame) and shows the statistical significance of autoregressive feature and the presence of their heavy-tailedness in extreme cyber losses fitted by the Fréchet type of generalized extreme value distribution (GEV).

Baldwin et al. (2017) utilize the Hawkes process to investigate the contagion effect of cyber risks and identify the strong evidence of contagion between cyberattacks. Aldasoro et al. (2022) disclose the heterogeneity of cyber risks concerning firm-specific information, such as firm size, industry, and firm's exposure to digital and cloud services. Notably, they ascertain that an increased level of exposure to digital and cloud infrastructures is linked to a mitigating impact on costs, which is further contingent on the size of the firm. However, they also caution that the reliance on cloud services has the potential to amplify tail-risks.

[7]Specifically, compared with existing literature, we use a much larger and much more comprehensive dataset (not only data breaches, but different types of cyber events, 85,085 loss events, time frame from 2005 to 2020). Another innovative element with respect to the dataset is to link this large dataset with another large dataset on socioeconomic information, providing the first U.S. study on drivers of spatial clusters (to our knowledge only one study for China exists so far). Two earlier studies for the United States exist (Khey and Sainato 2013; Ma et al. 2022), but they do not consider socioeconomic spatial impacts on cyber losses, look at only a few thousand data breaches, and are limited to five years. To our knowledge only one study (Chen et al. 2021) looks at spatial drivers for cyber risk, in their case for China but not for other parts of the world, and again only for a few thousand cases in a specific cyber loss category (i.e., fraud).

[8]Most literature on cyber risk or data breach risk analysis has investigated aggregate loss events across a nation or among nations (see, e.g., Edwards et al. 2016; Eling and Jung 2018). These studies can provide insights for how potential aggregate loss amount can be managed on a national or societal level, but are not fully able to offer a regional-level solution on risk management. In this sense, our study contributes to the literature by elaborating on county-level spatial risk models.

2.2. Spatial/Socioeconomic Analysis on Cyber Risks

The number of studies focusing on the spatial analysis of cyber risk and cross-sectional socioeconomic factors is very limited. Khey and Sainato (2013) examine how U.S. data breaches between 2005 and 2012 are spatially correlated using the Privacy Rights Clearinghouse (PRC) dataset that provides information on location of events. The authors employ Moran's *I* statistics to model potential clusters and spatial outliers of data breach events and find no significant pattern of cyber victimization in industry, geographical region, or time frame. They conclude that no clear pattern of data breach criminal behavior is identified for the data period. However, this study is limited to a simple presentation of spatial dependency in the United States with a short data period. Given that no results of spatial dependency were found, the study also does not investigate what can drive such spatial dependency.

Sen and Borle (2015) take advantage of criminological theories to address how malicious cyber attackers may behave to cause data breaches. For example, they argue that opportunity theory of crime can help understand cyber risk landscape, where software or system vulnerabilities can provide outside hackers with opportunities to cause significant data breach events. The authors find results in their empirical study supporting that the larger the number of customers an outsourced security service provider has, the greater interdependency risk the system can have; hence overall social welfare may decline. In addition, they claim that organizations need to assess data breach risk environment based on locations and industry sectors to find an optimal strategy for cyber risk management (e.g., information technology [IT] investments).

Amin et al. (2021) implement spatial analysis of the country of origin for cyberattacks to detect potential spatial clusters of countries. Using the cyberattack data from the Canadian Institute for Cybersecurity with more than 110,000 uniform resource locators (URLs), the authors show the geographical representation of cyberattacks and hotspots for countries of origin and determine the spatial clusters of areas with high frequency or cyberattack rates by type (benign, phishing, malware, spam, and defacement). Although this study provides worldwide-level spatial dependency to figure out the cyberattack landscape, it does not offer socioeconomic implications for what can drive spatial features and countries of origin.

Chen et al. (2021) address this limitation using cyber fraud events in China. The authors employ a generalized additive model (GAM) to examine potential causes of cyber events and find that regional economic development and population structure affect the spatial distribution of cyber fraud events. This finding indicates that cyber events tend to occur in more developed, coastal areas, featuring higher income and a well-constructed network infrastructure. Our study is associated with Chen et al. (2021) in that both investigate spatial dependency in cyber risk events and features of some socioeconomic factors; however, our spatial econometric approaches with the U.S. dataset can provide implications for how socioeconomic factors can address cyber loss frequency and severity in adjacent counties, thereby helping to better understand the socioeconomic features of spatial clusters.

Ma et al. (2022) investigate the PRC data to model the relationship between data breach loss frequency and severity. They employ the autoregressive conditional duration model (ACD) and ARMA-GARCH model to examine the statistical features of the loss process. They implement the K-means cluster approach and find the independent relationship between loss frequency and severity within a cluster. Our study is in line with Ma et al. (2022) in that potential clusters of cyber loss events with respect to frequency and severity are determined but complements this study by providing spatial clusters and their socioeconomic determinants. Table 1 provides sectional comparisons between the present study and six other relevant studies on cyber risk and its spatial analysis.

2.3. Discussion of Potential Key Socioeconomic Factors

Based on our review of the literature, we find that some socioeconomic factors may spatially affect cyber loss frequency and severity, which might be important to understanding the risk exposure of a particular region. A region could be a single district or a spatial cluster (pattern) of multiple neighboring counties featuring some common factors (e.g., economic environment, industrial structure). In the literature, the size of exposures to cyberattacks can be represented by, for instance, the number of computers in the network environment, the development of Internet infrastructure and the size of population.[9] The income level could also be associated with cyber risk event patterns. For example, Gorman and Malecki (2000) and Rutherford et al. (2004) argue that Internet infrastructure is better developed in metropolitan areas, where average income is relatively higher than in nonmetropolitan areas. Considering potential geographical association between the size of exposures and cyber risks, we examine the spatial impacts of the population size and income level on cyber loss frequency and severity.

[9]For example, Sassen (2001) and Castells (2004) argue that Internet infrastructure is more developed in bigger (or mega) cities with highly connected telecommunications and more telecommunicators. Aldasoro et al. (2022) and Eling, Jung, and Shim (2022) use the number of employees to represent the level of cyber incident costs, implying that more exposures to cyber risk events can appear by the size of population. Biener, Eling, and Wirfs (2015) also support this idea by showing that firms with more employees tend to have more cyber losses.

TABLE 1
Comparison of Relevant Literature with the Present Study in Cyber Risk and Spatial Analysis

	Aspect 1: General cyber loss analysis			Aspect 2: Spatial/socioeconomic analysis on cyber risks			
	Eling and Loperfido (2017)	Eling and Wirfs (2019)	Jung (2021)	Khey and Sainato (2013)	Chen et al. (2021)	Ma et al. (2022)	Present study
Data	PRC (2005–2015)	SAS OpRisk (1995–2014)	• Cowbell Cyber(2005–2018) • PRC (2005–2018)	PRC (2005–2012)	China Judgements Online database (2014–2018)	PRC (2012–2018)	• Advisen(2005–2020) • Social Determinants of Health Database (2011–2020)
Sample size	2,266	1,579	21,555	3,226	6,106	6,894	• 85,085 (Advisen) • 32,245 (SDOH)
Method	• Multi-dimensional scaling • Multiple factor analysis for contingency tables • Goodness-of-fit	• Loss distribution approach • Dynamic extreme value theory	• Generalized extreme value distribution • Time-series analysis	Moran's *I* statistics	• Moran's *I* statistics • Generalized additive model	• Auto-regressive conditional duration models • ARMA-GARCH model • K-means clustering	• Moran's *I* statistics • Spatial lag/error models
Focus of study	Distribution fitting of cyber risk and risk measurement	Distribution fitting of cyber risk and firm-specific characteristics for extremes	Statistical features of extreme cyber losses	Spatial cluster analysis of data breaches (no analysis of drivers)	Spatial pattern of cyber frauds in China and drivers	Spatial cluster analysis of data breaches (no analysis of drivers)	Comprehensive spatial analysis of cyber risk events (cluster and drivers)
Main findings	• Clusters exist by types of data breaches • Skew-normal distribution is optimal for cyber severity	• Log-normal distribution might overestimate cyber losses • The larger the firm size, the more exposed to extreme losses	• Threshold-based estimation might underestimate extreme losses • The cost of a smaller breach is larger than the cost of larger breach	Breaches tend to occur within particular geo-clusters	• The distribution of cyber fraud events is affected by the regional economy and population	• Spatial and industrial information is reflected in clusters • Data breach frequency and severity are independent within a cluster	• Spatial dependency exists at county level • Population and income are generally related with cyber risks

Note: PRC is a United States-based data provider collecting data breach events. SAS is a United States-based software firm also collecting data on operational risk. Cowbell Cyber is a United States-based cyber insurer providing a comprehensive dataset for cyber loss events. The China Judgements Online database provides information on cyber fraud events occurring in China. Advisen provides a dataset associated with cyber risk events worldwide. SDOH is provided by the Agency for Healthcare Research and Quality and provides data on the social context, economic context, education, physical infrastructure, and healthcare context.

In addition to potential, spatial impacts of socioeconomic factors for the size of exposures, the industrial structure could make a spatial impact on the cyber risk landscape. Some industries may be more likely to be exposed to cyber risks. For example, the financial industry and public administrative sector are known as fields more likely to handle sensitive personal information (e.g., health or financial information); hence, malicious breaches may target such industry to exploit personal information for financial gain. In contrast, other types of industry, for example, retail or wholesale, might be relatively less exposed to malicious cyber risk attacks, if data are less sensitive. Considering the percentage of people employed in different industries, we investigate how geographical patterns of cyber loss frequency and severity appear depending on the industrial structure.

3. METHODOLOGICAL BACKGROUND AND MODEL SETUP

3.1. Spatial Dependence Measure: Spatial Weight Matrix and Moran's *I*

Contiguity in a spatial analysis represents spatial connectivity that is determined by geographical fragmentation (Shirabe 2005). The contiguity between two regions can be addressed using a spatial weight matrix. Weights between adjacent regions are assigned to construct a matrix, W, with N number of regions in a spatial $N \times N$ square matrix including zeros on its diagonal. We then identify which regions are contiguous to build the weight matrix. If region j is adjacent to region i ($i \neq j$), w_{ij} has nonzero value, and 0 otherwise (Fischer and Wang 2011, Chapter 2).

Three approaches to define spatial continuity are rook, bishop, and queen. These approaches align with corresponding chess pieces to determine which neighbors are considered for the matrix construction (Darmofal 2015, Chapter 2). Figure 1 shows the blue-colored cells, where each approach is defined as neighbors of the centered region of interest. The rook approach refers to regions with common boundaries as contiguous neighbors; the bishop approach defines contiguity by common vertices. The queen approach integrates both the rook and the bishop approaches (common boundaries and vertices) to define contiguous neighbors. The queen approach tends to be favored because it helps one appropriately model contiguity with irregular lattice data and provides a stronger theoretical justification than other approaches (Darmofal 2015, 16). Once adjacent regions in the dataset are determined, we normalize the sum of elements in every row of the matrix (Fischer and Wang 2011, Chapter 2).

A prespecified weight matrix can be used to derive a spatial autocorrelation measure, Moran's I statistic, which shows how two neighboring regions are spatially correlated. The statistic can be tested with the null hypothesis of spatial independence (or randomness) (Fischer and Wang 2011, Chapter 2). This measure indicates a linear dependency in a random variable between neighboring regions, representing whether a spatially dependence exists in the data.[10] Hence, it can be used as a diagnostic tool to examine whether advanced spatial econometric models can proceed (Darmofal 2015, 43). Two Moran's I estimators are used. One is the estimator to measure a global-level spatial dependency across the wide range of geographical regions (i.e., data as a whole). This estimator, called global Moran's I, is calculated by the deviations between the values of interest and their mean (Anselin 1995; Darmofal 2015, Chapter 4):

$$\text{Global Moran's I } (GI) = \frac{N}{S} \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i}^{N} (y_i - \bar{y})^2}, \tag{1}$$

where w_{ij} is an element of the $N \times N$ weight matrix W with N as the number of regions, S is the sum of the weights, and if the weight matrix W is row-standardized, then $N = S$; y_i and y_j represent regional cyber loss frequency or aggregate severity at location i and j, respectively. The value of a global Moran's I ranges between –1 and 1, where a value closer to 1 (–1) implies stronger similarity (dissimilarity) in values between regions (i.e., positive (negative) spatial autocorrelation). If the value is closer to 0, it implies no spatial autocorrelation between regions.[11]

Although a global Moran's I close to 0 represents no statistical spatial dependency in the overall dataset (large-scale geographical regions), it may not imply the absence of any spatial dependency (Darmofal 2015, 49). Spatial dependency might be observed at the local level (i.e., smaller-scale geographical regions). A local-level spatial dependency can be represented by the local Moran's I statistic for a certain region i defined as

[10]Global Moran's I is like a linear correlation, indicating whether there is spatial dependence in observations without considering the impact of other factors (Li et al. 2007). Positive global Moran's I indicates a positive linear correlation in loss frequency and severity between regions in our case, but it does not address which regions with particular socioeconomic features tend to show high or low risk exposures.

[11]The null hypothesis of Moran's I in our model posits spatial randomness, where in our case loss features (i.e., loss frequency and severity) are randomly distributed across the regions of interest (either county or state level).

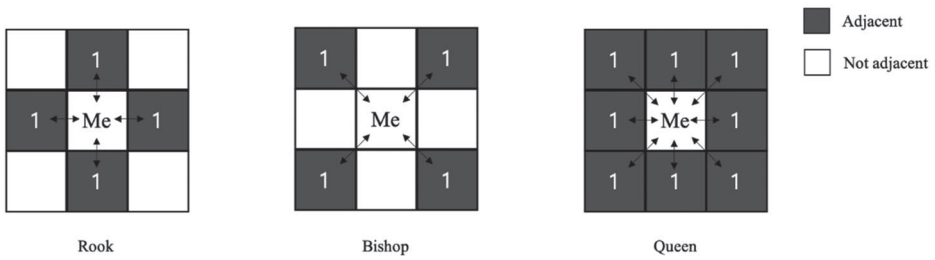

FIGURE 1. Three Definitions of Spatial Contiguity: Rook, Bishop, and Queen. *Note:* The blue-colored cells (regions) indicate neighbors of the centered cell (region) to construct a spatial weight matrix.

$$\text{Local Moran's } I_i(LI_i) = \sum_{j}^{J_i} w_{ij}(y_i - \bar{y})(y_j - \bar{y}), \tag{2}$$

where J_i is the neighborhood set of areas.

The local Moran's I indicator can be calculated for every region in the dataset. The global Moran's I indicator can be represented by the sum of the local Moran's I indicators of all regions constituting the entire geographical area. The relationship between two measures is

$$\sum_{i}^{N} LI_i = \gamma \times GI, \tag{3}$$

where γ is the scale parameter for the relationship between Moran's I indicators defined as $\frac{S}{N}(\sum_i^N (y_i - \bar{y})^2)$. Local Moran's I statistics can help one assess risk clusters that are relatively more exposed (hot spot) or less (cold spot) to risk events.[12]

3.2. Analysis of Socioeconomic Effects

In the last step, we investigate spatial impacts of socioeconomic factors on cyber loss exposures of the regions within the clusters. Specifically, we employ spatial panel regression models to examine the spatial impacts on frequency and severity of cyber loss events at the county level. Spatial regression approaches are designed to model the impacts of explanatory variables featured in adjacent areas of the region, where the response (i.e., loss frequency and severity in our case) is observed. The rationale of using a spatial model is that the observations are located in different regions, and they can be spatially affected by relevant factors of neighbors (Okunlola et al. 2021). This potential influence of neighbors can lead to spatial diffusion that may worsen the performance of the classical linear model.[13]

To pinpoint the potential spatial impacts on cyber loss events, we consider three approaches: the spatial autoregressive model (SAR; also called the spatial lag model), the spatial error model (SEM), and the spatial autoregressive combined model (SAC). In Section 5 we compare all three models to conclude which describes our data best. The starting point to describe the spatial panel models is the standard linear model without spatial weights (SLM) as a benchmark. The model with the variables of interest can be defined as follows:

$$\begin{aligned} y_{it} = {} & \beta_0 + \beta_1 * Pop_{it} + \beta_2 * Inc_{it} + \beta_3 * WS_{it} + \beta_4 * Ret_{it} \\ & + \beta_5 * Fin_{it} + \beta_6 * Edu_{it} + \beta_7 * Admin_{it} + \beta_8 * Armed_{it} + \mu_i + \xi_t + \epsilon_{it}, \end{aligned} \tag{4}$$

where the explanatory variables are specified as from "Pop" (population) to "Armed" (armed forces industry; see Section 4 for the specification); μ_i and ξ_t represent fixed effects controlling for time-invariant and spatial-invariant variables, respectively (Elhorst 2014); and ϵ_{it} presents an error term that is an i.i.d. (independently and identically distributed) normal random variable with zero mean and variance σ_ϵ^2.

[12]A hot spot is defined by the region of interest and its neighbors whose values of the response variable are above than the global mean, whereas a cold spot is the area where values of the region of interest and its neighbors are all below the global mean.

[13]The standard linear regression model under the Gaussian error distribution is tractable and can provide optimal unbiased estimators. The model is based on conditional mean surfaces and interpreting the impacts of the explanatory variables is straightforward. However, the standard linear approach is clearly unable to address the relationship between variables if a spatial dependence structure exists in the error term.

While the SLM does not take into consideration any spatial effects from adjacent regions, the SAR can incorporate spatial correlation in the response variable, reflecting autoregressive effects between responses using a spatial weight matrix. Our SAR model with socioeconomic factors representing risk exposures and industrial features can be expressed as

$$\begin{aligned} y_{it} \quad &= \rho\sum_{k=1}^{N} w_{ik}y_{kt} + \beta_0 + \beta_1{*}Pop_{it} + \beta_2{*}Inc_{it} + \beta_3{*}WS_{it} + \beta_4{*}Ret_{it} + \beta_5{*}Fin_{it} + \beta_6{*}Edu_{it} \\ &\quad +\beta_7{*}Admin_{it} + \beta_8{*}Armed_{it} + \mu_i + \xi_t + \epsilon_{it}, \end{aligned} \tag{5}$$

where ρ is the coefficient of the spatial autocorrelated term indicating the strength of spatial autocorrelation. This spatial parameter (ρ) enables consideration of endogenous interaction effects, suggesting that the value of the dependent variable in a specific region is influenced by those in neighboring regions (Elhorst 2013).[14] Here μ_i and ξ_t again present fixed effects controlling for time-invariant and spatial-invariant variables, respectively.[15]

The SEM can help one capture a latent structure in the error term of the linear model (Fischer and Wang 2011, Chapter 3). The latent structure indicates omitted latent variables that may be spatially correlated and the SEM takes into consideration the potential impacts of those variables in the error term (Voss et al. 2006). Unlike the SAR model, the SEM provides the identical expectation of the response variable as that of the classical linear model (Fischer and Wang 2011, Chapter 3). It implies that the point estimates of the model coefficients can be identical between the SEM and the classical model for large samples. This can help interpret the SEM results in the way that the results of the classical linear model are interpreted. Our SEM model can be defined as

$$\begin{aligned} &y_{it} = \beta_0 + \beta_1{*}Pop_{it} + \beta_2{*}Inc_{it} + \beta_3{*}WS_{it} + \beta_4{*}Ret_{it} + \beta_5{*}Fin_{it} + \beta_6{*}Edu_{it} + \beta_7{*}Admin_{it} + \beta_8{*}Armed_{it} + \mu_i + \xi_t + u_{it}, \\ &u_{it} = \lambda\sum_{j=1}^{N} w_{ij}u_{it} + \epsilon_{it}, \end{aligned} \tag{6}$$

where λ is a parameter to model spatial dependency in the error term, u_{it}, and thus the SEM can represent interaction effects within the error term.

The SAC approach proposed by Kelejian and Prucha (1998) offers both endogenous and error interaction effects by incorporating the spatial autocorrelation in the response variable and the spatial correlation with latent factors (Elhorst 2014, Chapter 3). Providing this functionality, the SAC might be able to better explain the model fit for the spatial relationship than the other two approaches (SAR and SEM). Our SAC model can be formulated as

$$\begin{aligned} &y_{it} = \rho\sum_{k=1}^{m} w_{ik}y_{kt} + \beta_0 + \beta_1 * Pop_{it} + \beta_2 * Inc_{it} + \beta_3 * WS_{it} + \beta_4 * Ret_{it} + \beta_5 * Fin_{it} + \beta_6 * Edu_{it} + \beta_7 * Admin_{it} + \beta_8 * Armed_{it} + \mu_i + \xi_t + u_{it} \\ &u_{it} = \lambda\sum_{j=1}^{N} m_{ij}u_{it} + \epsilon_{it} \end{aligned} \tag{7}$$

where m_{ij} is a component of a spatial weight matrix and may or may not be equal to w_{ij}.

[14]Endogenous interaction effects among dependent variables (i.e., loss frequency and severity) in our models can address potential relationship between the frequency or severity of cyber incidents in a region of interest and those of its neighboring areas. This relationship may arise from the contagious nature of cyber risk.

[15]To estimate fixed effects of the spatial panel regression, we implement the maximum likelihood estimation, where the loglikelihood can be obtained from the following formula (spatial lag model):

$$y_{it} = \rho\sum_{j=1}^{N} w_{ij}y_{jt}\boldsymbol{x_{it}\beta} + \mu_i + \epsilon_{it}.$$

The loglikelihood function can be defined as

$$\mathcal{L} = -\frac{NT}{2}\log(2\pi\sigma^2) + T\log|\boldsymbol{I_N} - \rho\boldsymbol{W}| - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(y_{it} - \rho\sum_{j=1}^{N} w_{ij}y_{jt} - \boldsymbol{x_{it}\beta} - \mu_i\right)^2.$$

The estimated fixed effects to maximize the likelihood of the above equation can be derived, for example, with respect to time-invariant effects:

$$\mu_i = \frac{1}{T}\sum_{t=1}^{T}(y_{it} - \rho\sum_{j=1}^{N} w_{ij}y_{jt} - \boldsymbol{x_{it}\beta}).$$

The spatial-invariant effects can be obtained by taking into account N number of observations (i.e., the number of regions considered) to average the residuals of this equation.

4. DATA DESCRIPTION

We collect the data from two sources to investigate spatial dependency in cyber risk events and the impact of socioeconomic factors. We use the largest cyber loss database by Advisen[16] containing loss event information from multiple publicly available sources. It contains more than 110,000 cyber loss events worldwide, including information on event characteristics (e.g., event date, event type,[17] geographical information,[18] the cause of an event, total count of affected records or the total amount of financial loss), and firm-specific information (e.g., industry code, legal liability, or firm size). We consider the sample from 2005 to 2020 and only U.S. loss events,[19] leading to a total sample size of 85,085.

To model cyber loss frequency in both spatial autocorrelation and regressions, we consider all recorded events of this dataset on an annual basis. For cyber loss severity, we use the affected count variable to build a county-level response variable,[20] a continuous variable representing the number of records lost in a cyber incident. Cyber incidents constituting this variable include 15 risk types that include, for example, malicious breach, cyber extortion, unintentional disclosure, network disruption, or IT processing errors; hence these can represent a general landscape of cyber risks. We further use the variable for geographical locations in the spatial autocorrelation analysis, representing the ZIP code of the victim organization.[21]

We further collect data from the Social Determinants of Health (SDOH) database offered by the Agency for Healthcare Research and Quality (AHRQ). This provides county-level information between 2009 and 2020, including social (e.g., population, age, or race/ethnicity) and economic context (e.g., income or unemployment rate). We use the following information: (1) county-level population size as a proxy for the level of Internet usage; (2) county-level average income representing the wealth level of a county; and (3) county-level industry ratio (percentage of people working in wholesale, retail, finance, education, administration, armed forces).[22]

One can expect that the wider the Internet connection in a county is, the more likely a cyber risk event is to occur as more exposures to such risk are present. To address this expectation, we use the population size of a county as a proxy for the level of Internet infrastructure, implying that the more people reside in a county, the higher is the level of Internet infrastructure. This proxy can be supported by the literature showing that the Internet infrastructure tends to be much invested and concentrated in metropolitan areas with large populations (see, e.g., Sassen 2001; Tranos 2013). In terms of the relationship between the wealth level and cyber risk events, it is commonly identified that the income level or economic development of a county is positively associated with the number of cybercrimes (Park et al. 2019; Chen et al. 2021). We thus use the average income to examine the impact of the wealth level of a county on cyber loss occurrence in adjacent counties. Lastly, the industry ratio variables are used to represent types of industry exposed to cyber risks.

Table 2 shows the description on socioeconomic factors of interest (panel A) and descriptive statistics of cyber loss frequency and severity and socioeconomic factors at the county level (panel B). It is observed that there are deviations in the number of cyber loss events and industry ratios over counties, and cyber loss events have occurred on average twice per year at the county level. With respect to the industry ratio, the education and retail industries top the list of industries, followed by the administrative and financial sectors.

[16]This dataset has been used in the cyber risk literature, mainly to examine the statistical features of cyber loss frequency and severity and cross-sectional determinants on cyber risk events (see, e.g., Romanosky 2016; Palsson, Gudmundsson, and Shetty 2020; Shevchenko et al. 2021; Aldasoro et al. 2022; Malavasi et al. 2022).

[17]In Section 5.2, we implement spatial panel models by categorizing the event types into three classes: malicious, unintended, and privacy risk types. The definition of the risk classification can be found in Appendix A.

[18]The geographical information in our dataset is the ZIP code of an affected organization, where the event may have taken place in the headquarter or in its branch of other states.

[19]We consider the data up to 2020, because of the availability of the socioeconomic information. The Federal Information Processing System (FIPS) state code is used to merge the cyber and socioeconomic data.

[20]We consider this variable to represent cyber loss severity. It provides a large set of nonzero loss observations showing 71,550 U.S. cases for our data period. Another information to represent loss severity can be the total cost amount accounting for the final total cost by a cyber incident. However, this variable only provides 719 nonzero loss events, which can be limited to build a county-level variable. Thus, we stick to the larger sample that can lead us to address cyber risk landscape more accurately in the spatial manner.

[21]Spatial features of individual victims of cyber loss events may be interesting information as well, but this is not collected in our dataset. However, individual owners of breached information may not be key agents for cyber risk management that is at the core of discussion in this study. In addition, from the insurer's point of view, customers in the cyber insurance markets are organizations whose spatial features can be used as loss drivers to price cyber risk. Nevertheless, individual information on breached data might be relevant to understand spatial features of cyber risk under the assumption that geo-locations of breached individuals are clustered.

[22]We select the industry variables to be aligned with industrial types of the PRC dataset (Appendix C). The database integrates education, healthcare, and social assistance into a single code (i.e., education type).

TABLE 2
Socioeconomic Variables and Descriptive Statistics

A: Variable description

Population	Population of a county
Income ($)	Per capita income (in dollars, inflation-adjusted to file data each year)
Wholesale (%)	Percentage of the employed working in wholesale trade
Retail (%)	Percentage of the employed working in retail trade
Finance (%)	Percentage of the employed working in finance and insurance, real estate, and rental and leasing
Education (%)	Percentage of the employed working in educational services, and healthcare and social assistance
Administrative (%)	Percentage of the employed working in public administration
Armed forces (%)	Percentage of the employed working in armed forces

B: Summary statistics (county-level)

	Mean	Std	Minimum	Median	Maximum
Loss frequency	2.21	13.26	0.00	0.00	446.00
Loss severity	1,076,208.27	39,542,898.74	0.00	0.00	4,826,627,241.00
Population	100,790.47	323,241.31	62.00	25,769.00	10,105,722.00
Income ($)	25,343.42	6,380.65	7,887.00	24,434.00	78,771.00
Wholesale (%)	2.40	1.23	0.00	2.31	36.36
Retail (%)	11.35	2.42	0.00	11.45	32.45
Finance (%)	4.61	1.93	0.00	4.31	22.82
Education (%)	23.09	4.67	0.00	22.74	57.95
Administrative (%)	5.68	3.25	0.00	4.81	35.36
Armed forces (%)	0.30	1.55	0.00	0.04	52.64

Note: Breach frequency and severity statistics are based on the Advisen dataset, whereas the statistics of the other variables are derived from SDOH and AHRQ datasets. Breach severity indicates the sum of the affected counts involved in events at county level. The summary statistics in Panel B are derived from the panel dataset used in the spatial regression in Section 5, which includes 3,132 counties over the span of 10 years (2011–2020). The percentage of zeros for loss frequency and severity can be large at county level for the entire sample as found in panel B with more than half of zeros; however, it is rather small for key regions of this study (i.e., California and the Northeast Coast region). We test each of the 50 U.S. states and the Northeast Coast region (New York, New Jersey, Connecticut, Massachusetts) to see whether many zeros may lead to statistical significance of spatial correlations and find no such tendency. The test results are available upon request.

5. RESULTS

5.1. Regional Spatial Autocorrelation

We use GeoDa to analyze spatial dependency in cyber loss frequency and severity at state and county levels, showing the potential presence of geographical clusters.[23] In Table 3 we present the global Moran's I statistics. Following Jung's (2021) argument that there is a structural break in cyber loss severity in the year 2014, we consider three time periods: (1) the entire data period; (2) the pre-2014 period; and (3) the post-2014 period (including 2014).[24] There is no spatial dependency across

[23]GeoDa is a free and open-source software package released in 2003 and widely used to explore spatial patterns in various contexts, for example, housing prices and homicide rates (Anselin and Rey 2014). The economics literature has also used this software to analyze geographical property prices (Cotteleer and van Kooten 2012) or regional economic growth (Ramajo et al. 2008). GeoDa incorporates a nonparametric permutation test to examine the null hypothesis of spatial randomness. The distribution of the Moran's I statistic under the null hypothesis was derived in our model under the assumption of spatial randomness by randomly permuting the observed values across regions of interest (Anselin et al. 2006). We employ the software to estimate global and local Moran's I statistics. For the spatial panel regressions, we use the splm package of the R statistical program.

[24]We use the queen contiguity to define the geographical spatial weight matrix for two reasons. First, this definition helps avoid overrejecting the null hypothesis (Darmofal 2015, 72). Second, it is flexible for any irregular lattice data by covering all neighbor regions (Darmofal 2015, 16). We also examine spatial dependency of loss frequency for three time periods as in the loss severity case for the sake of comparison.

TABLE 3
State- and County-Level Global Moran's *I* Statistics

	Loss frequency		Loss severity	
Year	State-level	County-level	State-level	County-level
Before 2014	–0.026	0.270***	–0.053	0.102***
After 2014	–0.046	0.226***	–0.031	0.189***
Entire period	–0.040	0.245***	–0.038	0.174***

Note: The figures indicate Moran's *I* statistics at the state and county-level; *, **, and *** indicate significance level of 10%, 5%, and 1%. We use the pseudo *p* value with GeoDa to test the statistics.

states over the period, but significant dependence across counties at the 1% level for both loss frequency and severity.[25] This result is visualized in Figure 2, where the fitted line in the left plot representing the state-level is nearly flat, whereas that of the right plot representing the county-level is linearly on the rise for both frequency and severity.

The county-level results with respect to frequency and severity are all positive. This demonstrates that relatively more exposed counties (i.e., counties with a relatively large number of events and large losses) and less exposed counties (i.e., counties with a relatively small number of events and small losses) are clustered. We conclude that cyber risk events are geographically clustered in particular locations, where certain socioeconomic environments may motivate attackers or there could be relatively more risk exposures (e.g., large number of employees or more interconnected industries), potentially causing cyber risks.

For the cases that turn out to be statistically significant in the global spatial autocorrelation (i.e., loss frequency and severity at county level), we then investigate the local Moran's *I* statistics to further assess spatial dependency for every county and to graphically identify regions that are relatively more exposed (hot regions) or less exposed (cold regions) to cyber risks.[26] Figure 3 illustrates graphical results of local Moran's *I* statistics with five categories of cyber loss frequency and severity that present statistically significant spatial autocorrelation.

Specifically, the high–high regions describe that counties where the centered county and its neighbors are all exposed to many cyber risk events. The high–low category demonstrates that the centered county is highly exposed to cyber risks, whereas its neighbors face a small number of events. The low–high category represents areas where the centered county is less exposed to cyber risks, whereas its neighbors face many events. The low–low category stands for the opposite: Both are less exposed to cyber risks. The approach to formulating hot and cold areas for spatial dependency is implemented as follows:

1. We normalize the event frequency (or severity) of all counties, then map values of every single county on the x-axis and the average of values of its neighbors on the y-axis.
2. We calculate local Moran's *I* with normalized observations and test statistical significance of local Moran's *I* at the 5% confidence level[27] to reject the null hypotheses (representing the presence of spatial dependency).
3. We finally project the counties filtered from the previous step into the map to identify hot and cold clusters for cyber loss events.

Following this approach, we see that West and Northeast Coast regions (especially counties in California and on the Northeast Coast) are more exposed to cyber risks, particularly with respect to loss frequency. Those counties are mostly categorized as high-high clusters, and thus counties in those areas are in general highly exposed to cyber risks. Many regions in the middle turn out to be the low–low areas, where counties tend to be less exposed to cyber risks than others.

[25]This finding is in line with existing literature (Khey and Sainato 2013) for shorter time periods. Note that we take the summation of loss amounts to derive the loss severity at each regional level (i.e., either state or county) for this analysis, which can provide an aggregate size of cyber risk damage in each region. If one examines global Moran's *I* statistics for each state with its counties, not all states would have statistically significant Moran's *I*, which we observe by implementing a test with our data. This may result from the fact that the global Moran's *I* is derived with the average of the values of interest (i.e., loss frequency and severity in our case), which can differ between state and county levels. The results are available upon request.

[26]We present only the graphical results of local Moran's *I* in Figure 3 instead of providing numerical results because of the large number of counties considered. The detailed results are available upon request.

[27]We follow Rusche (2010) and Darmofal (2015, 57) to use 5% as the rejection criterion for the null hypotheses.

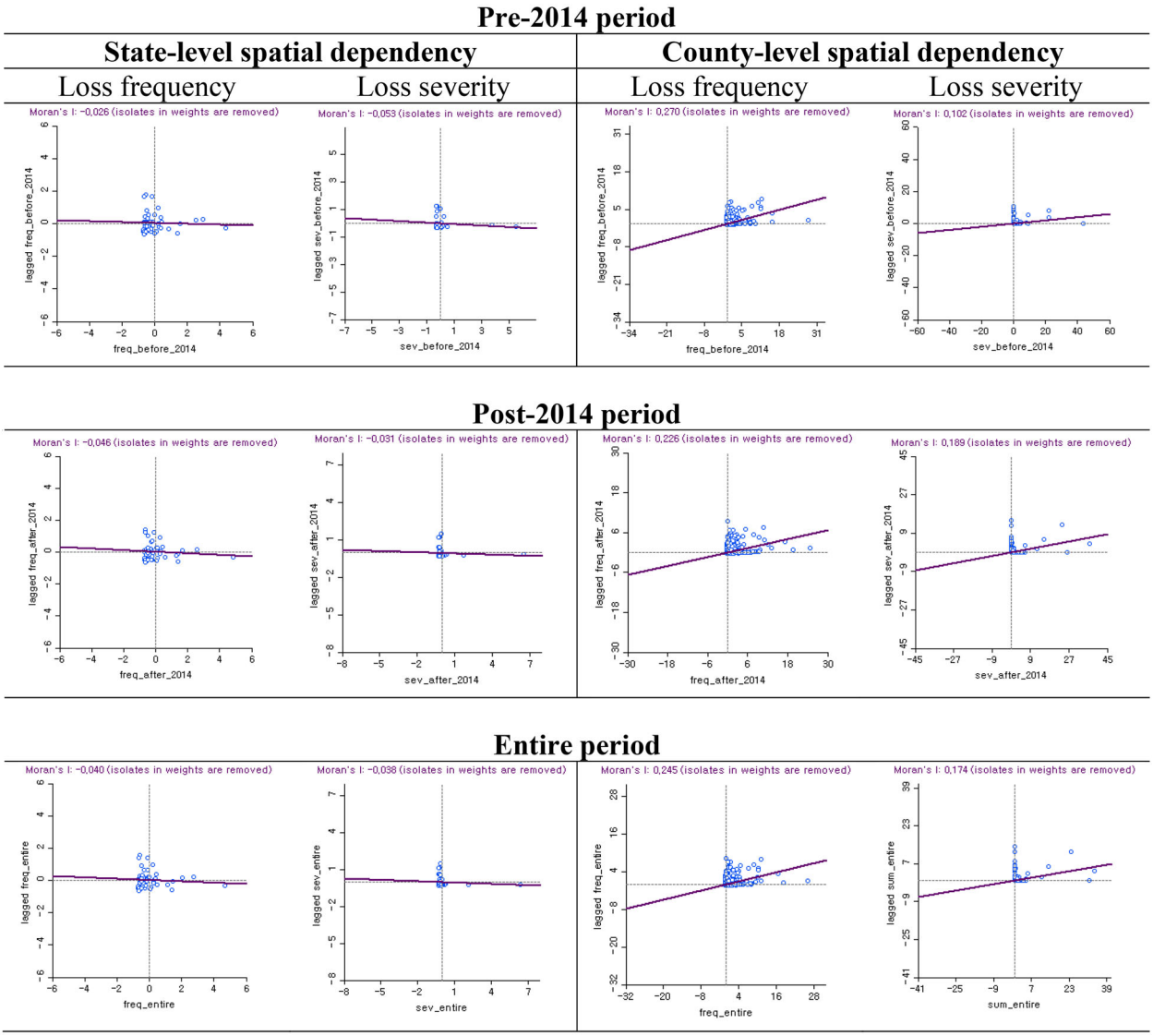

FIGURE 2. Graphical Results of the Global Moran's *I* Statistics for Cyber Loss Frequency and Severity at both State and County-Levels.

5.2. Spatial Analysis on the Impacts of Socioeconomic Factors

5.2.1. Analysis on All Counties

In the previous section, we documented spatial dependency at the county level and hot/cold clusters of cyber loss frequency and severity. This finding raises the question of what socioeconomic characteristics may affect spatial autocorrelation and the spatial clusters. To address this, we examine what socioeconomic factors may explain spatial clusters of cyber loss exposures. To do so, we employ three spatial panel regressions, namely, the spatial lag model (SAR), spatial error model (SEM), and spatial autoregressive combined model (SAC), and the nonspatial standard linear regression model (SLM) as a benchmark. Combining the cyber loss event data and county-level socioeconomic data, we consider the period between 2011 and 2020 in 3132 counties of 50 U.S. states for this study.[28]

Panel A of Table 4 presents the results of four models considering all counties of the dataset with respect to loss frequency and severity. We determine that the SAC provides a better fit with the lowest AIC to address the spatial effects of

[28]This data period is determined by the fact that the socioeconomic dataset is not fully available prior to the year 2009 and after 2020. In terms of which states are included, we do not consider Hawaii that is not geographically connected in the mainland and thus not clearly recognized as geographically adjacent regions. In addition, we remove several missing values in 2010 and 2018 that result from changes in administrative districts, for example, *Rio Arriba*. This exclusion leads us with a final dataset of 50 states and 3132 counties.

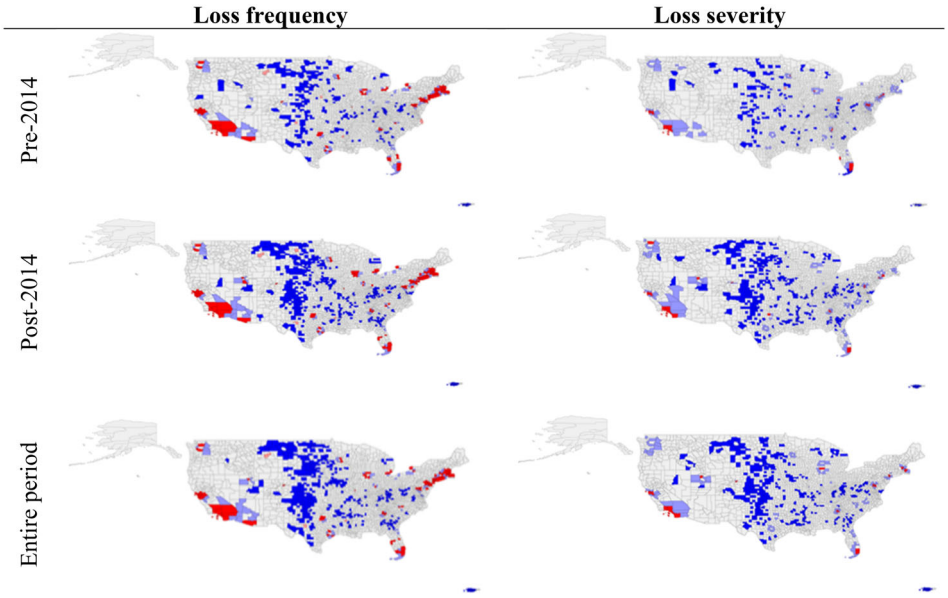

FIGURE 3. Results of Spatial Hot and Cold Clusters of Cyber Risks at the County Level. *Note:* The graphs of the left (right) panel illustrate spatial dependency based on cyber risk event frequency (severity). The categories in the legend indicate the following: high–high (red), the centered county and its neighbors all exposed to large number of cyber risk events; high–low (light red), the centered county highly exposed to cyber risk events, but its neighbors exposed to small number of events; low–high (light blue), the centered county less exposed to cyber risk events, but its neighbors exposed to large number of events; low–low (blue), the centered county and its neighbors all less exposed to cyber risk events; not significant, the counties that turn out to be insignificant in the statistical test.

socioeconomic factors on loss frequency and severity than the other models do.[29] We also find statistical significance of independent variables in explaining both frequency and severity at the county level. Specifically, population and average income level of a county are positive and significant at the 1% confidence level in the spatial models. Counties with a large population or that have high incomes are thus more likely to be exposed both in frequency and severity.

The results of both loss frequency and severity models provide evidence on significant spatial impacts of socioeconomic factors with statistical significance of the spatial autocorrelation parameter (ρ).[30] The presence of the spatial effect can help one disentangle the direct and indirect spatial effects[31] from the total effect of an explanatory variable, effects that are displayed in panel B of Table 4. It is observed that both population and average income level have positive direct effects, but negative indirect effects on both frequency and severity of counties. This result implies that a county with large population or

[29]Loglikelihood (LL) and AIC do not play a role in evaluating how different multiple models are in between, but how different their performances are. Golgher and Voss (2016) provide results of comparing spatial models and LLs are not significantly different as our results. However, spatial models and the SLM are clearly different in that spatial models incorporate spatial parameters to reflect spatial effects (see Figure 1 of Golgher and Voss [2016] for the difference). Our case focuses more on whether spatial effects are present and if so, what spatial model is better at addressing them and how effects may appear. To provide supporting evidence for performance measures (i.e., LL and AIC), we implement the log-likelihood ratio test (LRT) and a graphical diagnosis by comparing residual plots of the SLM and the optimal spatial model and find that statistical significance at the 1% confidence level on the differences in LL is obtained for all cases and the residuals of the optimal spatial models tend to be closer to zero than those of the SLM for both loss frequency and severity (see Tables E1 and E2 of Appendix E for the LRT results and Figures E1 to E4 for graphical diagnosis). These results can support that the optimal spatial model can better describe the relationships between variables of this study than the SLM.

[30]Note that statistical significance of error coefficients, for example, in the SEM (i.e., the value of λ) indicates the presence of spatial correlation among the errors (Golgher and Voss 2016). It highlights that the expected value of the dependent variable in the SEM remains consistent with that of the SLM model because both assume that the error term follows a normal distribution with the mean of zero. Thus, the interpretation of SEM aligns with that of SLM, indicating that the indirect effect is computed as zero (Golgher and Voss 2016).

[31]The direct effect indicates how a change in an explanatory variable can affect, on average, a change in the dependent variable in the identical region, whereas the indirect effect shows how a change in an explanatory variable of the region of interest can affect, on average, a change in the dependent variable of neighboring regions. The sum of these effects is the total effect. What is important for the interpretation of these effects is their signs that help one elaborate on what directions of spatial effects work. Brief explanation on the derivation and interpretation of the effects can be found in Appendix B. See LeSage and Pace (2009) and Golgher and Voss (2016) for more detail.

TABLE 4
Results of Spatial Panel Analysis on All Counties

A: Comparison of regression results

	Cyber loss frequency				Cyber loss severity			
	SLM	SAR	SEM	SAC	SLM	SAR	SEM	SAC
Constant	–71.052***	–53.077***	–66.152***	–61.803***	–35.295***	–32.872***	–34.374***	–35.636***
	(3.695)	(3.611)	(3.608)	(3.152)	(0.801)	(0.798)	(0.795)	(0.792)
ln(Population)	3.106***	2.615***	3.052***	2.888***	1.319***	1.252***	1.341***	1.401***
	(0.055)	(0.056)	(0.061)	(0.067)	(0.012)	(0.013)	(0.013)	(0.016)
ln(Income)	4.847***	3.419***	4.465***	4.407***	2.322***	2.125***	2.211***	2.299***
	(0.363)	(0.357)	(0.397)	(0.414)	(0.079)	(0.080)	(0.084)	(0.089)
Wholesale	–26.374***	–28.201***	–33.986***	–35.103***	–3.072**	–3.067**	–3.424***	–3.535***
	(5.731)	(5.602)	(5.852)	(5.311)	(1.243)	(1.239)	(1.270)	(1.281)
Retail	–66.738***	–58.200***	–60.725***	–51.482***	–13.006***	–12.172***	–12.278***	–12.560***
	(2.980)	(2.918)	(2.985)	(2.623)	(0.646)	(0.646)	(0.653)	(0.658)
Finance	70.326***	65.163***	63.706***	54.252***	17.467***	16.976***	16.154***	15.820***
	(4.325)	(4.228)	(4.478)	(4.159)	(0.938)	(0.935)	(0.966)	(0.979)
Education	–11.788***	–9.854***	–13.824***	–16.834***	1.329***	1.623***	1.310***	1.005***
	(1.499)	(1.469)	(1.569)	(1.486)	(0.325)	(0.325)	(0.337)	(0.346)
Administration	2.034	–1.660	–1.681	–5.044**	4.040***	3.572***	3.587***	3.691***
	(2.280)	(2.229)	(2.456)	(2.450)	(0.494)	(0.494)	(0.522)	(0.537)
Armed forces	–26.616***	–20.754***	–22.622***	–18.483***	–6.076***	–5.400***	–5.827***	–6.211***
	(4.581)	(4.480)	(4.571)	(4.003)	(0.993)	(0.991)	(1.001)	(1.005)
ρ	–	0.234***	–	–0.750***	–	0.088***	–	–0.082***
		(0.008)		(0.013)		(0.007)		(0.013)
λ	–	–	0.251***	0.786***	–	–	0.153***	0.231***
			(0.008)	(0.006)			(0.009)	(0.014)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
LL	–122,044.500	–121,496.900	–121,490.500	–120,931.300	–74,166.270	–74,091.970	–74,014.330	–73,998.220
AIC	244,125.000	243,033.800	243,020.900	241,904.700	148,368.500	148,223.900	148,068.700	148,038.400
Observation	31,320	31,320	31,320	31,320	31,320	31,320	31,320	31,320

(*Continued*)

TABLE 4
(*Continued*)

B: Measures of spatial effects

	Direct effect	Indirect effect	Total effect	Direct effect	Indirect effect	Total effect
ln(Population)	3.155	–1.505	1.650	1.403	–0.108	1.295
ln(Income)	4.814	–2.296	2.518	2.301	–0.177	2.124
Wholesale	–38.344	18.288	–20.056	–3.539	0.272	–3.267
Retail	–56.236	26.822	–29.414	–12.575	0.968	–11.607
Finance	59.262	–28.265	30.997	15.839	–1.219	14.620
Education	–18.389	8.771	–9.618	1.006	–0.077	0.928
Administration	–5.510	2.628	–2.882	3.696	–0.284	3.411
Armed forces	–20.190	9.630	–10.560	–6.218	0.479	–5.740

Note: We take the transformation of natural logarithm for two continuous variables (population and income). "FE" stands for the fixed effect; *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

high income is more likely to be exposed to cyber risk events itself; however, neighboring counties may have less likelihood of such events.[32]

Moreover, this suggests that a county with large population or high income tends to experience more severe losses from cyber incidents itself; however, the severity of such incidents in neighboring counties tends to be less severe. This finding can be partially supported by the literature claiming that wealthy countries with more Internet users tend to have higher rates of cybercrime activity (Kigerl 2012). Furthermore, as evidenced in the case of China as studied by Chen et al. (2021), economic development and population density can be of major drivers to understand spatial patterns of cyber loss events.[33]

We further identify in Table 4 that the variable of financial industry is positive and significant in explaining both cyber loss frequency and severity. In particular, direct and indirect effects of the financial industry variable demonstrate that counties with higher proportion of the financial industry tend to be more exposed to cyber risks, whereas neighboring counties tend to be less exposed to cyber risks with respect to both loss frequency and severity. In addition, the link between the size of the public sector and cyber loss severity is also positive and significant. This confirms the finding in Verizon (2019) that the public sector has been most exposed to cyber risks and the financial industry is one of the top three industries with most exposures. This report shows that cyberespionage is a major factor in the high exposure of the public sector and that the financial industry is primarily exposed to stolen credentials for use in banking applications.

When it comes to other industry variables, particularly loss frequency, the coefficients of the other industries (wholesale, retail, education, and armed forces) are negative and significant. All four of these variables have negative direct and positive indirect effects that are opposite to that of the financial industry variable. This result highlights that counties with high portions of the other four industries tend to be less exposed to cyber risks but may be at an increasing frequency of risk events in neighboring counties.

5.2.2. Analysis on Most Exposed Regions and Risk Types

We find in Figure 3 that the state of California and the Northeast Coast are hot clusters of cyber risk events, given that counties in such areas are in general exposed to cyber risks. We also see that the financial industries and public administration sector are concentrated in these regional clusters.[34] Moreover, those regions are known as densely populated and wealthy areas. The statistical finding in Figure 3 may thus raise a question of whether the socioeconomic features may also appear on the cyber loss frequency and severity for relatively exposed regions.

Tables 5 and 6 present the results of four models examining the socioeconomic features of cyber loss frequency and severity in hot clusters (the state of California in Table 5 and the Northeast Coast area[35] in Table 6). We find that for California, the SAC model is most appropriate to address spatial effects of socioeconomic factors on both loss frequency and severity. The Northeast Coast case illustrates the SAC model better accounting for the spatial effects of independent variables on loss frequency, but the SAR model as appropriate for loss severity. We also identify in both cases that the results for key socioeconomic variables (i.e., population and average income) are aligned with those of Table 4 by showing positive, significant impacts on loss frequency and severity and identical signs of direct and indirect spatial effects.

What makes the difference between these two cases is the impacts of the financial industry variable, where we find a negative impact in California, but the opposite on the Northeast Coast. This difference is observed particularly in loss frequency, implying that the financial industry of the Northeast Coast area tends to experience more cyber risks compared to the hot cluster of the West Coast. This finding can imply that although the financial industry is generally known as a major target of cyber attackers, its exposures to cyber risks may be heterogeneous for geographical hot clusters. Our result shows that the Northeast Coast regions with New York City may be more required to strengthen schemes for cyber risk management than the state of California with Silicon Valley.[36]

---

[32]More specifically, a 1% increase in log-transformed population of the county of interest can lead to a 3.155% increase in the number of events of that county, but around a 1.505% decrease in the number of events in neighboring counties. In a similar manner, a 1% increase in the log-transformed average income of the county of interest can lead to 4.814% increase in the number of events of that county, but around a 2.296% decrease in the number of events of neighboring counties.

[33]It is important to note that comparable with relevant literature in the cyber risk and criminological context, our study explores the association between cyber loss exposures and socioeconomic factors at the county level, but does not examine or claim causal effects of such factors on cyber losses.

[34]In Appendix D, we further examine two other states (Texas and Florida) to see how consistent their results are with the main results of two hot clusters.

[35]Based on the result in Figure 3, we include the states of Connecticut, New Jersey, New York, and Massachusetts to construct the Northeast Coast case for spatial panel analysis in Table 6.

[36]According to the U.S. Bureau of Economic Analysis, government, the state of California's agricultural, manufacturing, information and entertainment industries for 2021 account for higher portions of the regional gross domestic product (GDP) than for the national GDP, whereas the GDP portion of the financial industry is lower than that of the United States as a whole. In contrast, New York State, for example, shows lower portions of agricultural, manufacturing and entertainment industries than their portions of the national GDP, whereas the portion of its financial industry is almost double that of California (18.29% vs. 34.33%). This data can be accessed via https://www.bea.gov/data/gdp/gdp-state.

TABLE 5
Results of Spatial Panel Analysis in the State of California

A: Comparison of regression results

	Cyber loss frequency				Cyber loss severity			
	SLM	SAR	SEM	SAC	SLM	SAR	SEM	SAC
Constant	–461.178***	–408.838***	–561.243***	–683.451***	–83.145***	–80.314***	–82.657***	–66.412***
	(94.178)	(90.834)	(89.285)	(85.553)	(9.055)	(8.903)	(8.820)	(8.501)
ln(Population)	20.605***	19.269***	20.158***	19.350***	2.877***	2.831***	2.923***	2.602***
	(1.352)	(1.348)	(1.329)	(1.319)	(0.130)	(0.134)	(0.123)	(0.147)
ln(Income)	37.594***	33.627***	48.941***	61.458***	6.200***	5.944***	6.086***	4.607***
	(8.777)	(8.648)	(8.992)	(9.216)	(0.844)	(0.868)	(0.783)	(0.780)
Wholesale	–1,226.069***	–1,314.232***	–1,117.061***	–788.247***	–85.182***	–88.499***	–94.742***	–110.990***
	(230.525)	(223.703)	(221.834)	(225.902)	(22.165)	(22.137)	(21.341)	(20.204)
Retail	–343.032***	–345.631***	–372.588***	–376.110***	–42.568***	–42.299***	–42.763***	–40.906***
	(92.836)	(89.584)	(88.711)	(84.790)	(8.926)	(8.781)	(8.589)	(8.110)
Finance	–529.404***	–535.924***	–666.161***	–675.735***	–31.954***	–31.616***	–27.800***	–17.665*
	(116.931)	(113.259)	(113.033)	(110.065)	(11.243)	(11.101)	(10.759)	(10.274)
Education	–277.313***	–259.959***	–294.024***	–289.356***	–7.268	–6.585	–6.359	–0.091
	(50.861)	(49.527)	(48.174)	(45.860)	(4.890)	(4.871)	(4.729)	(4.702)
Administration	2.306	–5.701	–0.617	11.788	3.753	3.514	3.330	1.557
	(41.375)	(39.921)	(38.958)	(36.934)	(3.978)	(3.919)	(3.863)	(3.685)
Armed forces	–212.447	–194.663	–99.802	–35.160	–62.934***	–64.134***	–64.221***	–70.591***
	(193.243)	(186.442)	(182.491)	(172.084)	(18.580)	(18.308)	(17.972)	(17.019)
ρ	–	0.201***	–	–0.287***	–	0.045	–	0.227***
		(0.047)		(0.087)		(0.043)		(0.055)
λ	–	–	0.313***	0.557***	–	–	–0.193***	–0.454***
			(0.053)	(0.066)			(0.065)	(0.088)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
LL	–2,877.422	–2,868.213	–2,862.241	–2,859.337	–1,519.155	–1,518.615	–1,515.301	–1,507.178
AIC	5,790.844	5,776.427	5,764.481	5,760.675	3,074.310	3,077.230	3,070.603	3,056.356
Observation	580	580	580	580	580	580	580	580

(*Continued*)

TABLE 5
(*Continued*)

B: Measures of spatial effects

	Direct effect	Indirect effect	Total effect	Direct effect	Indirect effect	Total effect
ln(Population)	19.672	–4.631	15.041	2.634	0.734	3.368
ln(Income)	62.479	–14.709	47.771	4.663	1.300	5.962
Wholesale	–801.344	188.651	–612.693	–112.338	–31.312	–143.650
Retail	–382.359	90.014	–292.344	–41.403	–11.540	–52.943
Finance	–686.963	161.724	–525.239	–17.880	–4.984	–22.863
Education	–294.164	69.252	–224.912	–0.092	–0.026	–0.118
Administration	11.984	–2.821	9.163	1.576	0.439	2.015
Armed forces	–35.744	8.415	–27.330	–71.448	–19.915	–91.363

Note: We take the transformation of natural logarithm for two continuous variables (population and income). "FE" stands for the fixed effect; *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

TABLE 6
Results of Spatial Panel Analysis in the Northeast Coast Region (New York, New Jersey, Connecticut, and Massachusetts)

A: Comparison of regression results

	Cyber loss frequency				Cyber loss severity			
	SLM	SAR	SEM	SAC	SLM	SAR	SEM	SAC
Constant	–264.001***	–279.711***	–387.823***	–453.571***	–42.474***	–44.495***	–43.006***	–44.196***
	(73.882)	(73.017)	(71.968)	(61.475)	(7.309)	(7.209)	(7.206)	(7.208)
ln(Population)	8.214***	6.959***	6.140***	7.812***	2.818***	2.958***	2.797***	2.904***
	(1.258)	(1.301)	(1.261)	(1.177)	(0.124)	(0.133)	(0.121)	(0.146)
ln(Income)	25.959***	28.058***	38.829***	44.288***	1.808***	1.949***	1.823***	1.917***
	(7.004)	(6.930)	(7.190)	(6.636)	(0.693)	(0.685)	(0.660)	(0.678)
Wholesale	–1,560.270***	–1,515.301***	–1,414.167***	–949.302***	–49.820***	–48.651***	–50.587***	–49.676***
	(163.675)	(162.015)	(164.504)	(144.192)	(16.191)	(15.980)	(15.594)	(15.813)
Retail	–365.809***	–338.162***	–291.871***	–241.508***	–33.926***	–36.899***	–32.307***	–35.097***
	(77.368)	(76.727)	(76.803)	(66.889)	(7.653)	(7.654)	(7.444)	(7.731)
Finance	354.631***	363.973***	400.374***	343.165***	7.007	7.769	8.758	8.316
	(80.139)	(79.202)	(79.275)	(67.666)	(7.928)	(7.820)	(7.705)	(7.770)
Education	–95.737***	–86.737***	–97.306***	–149.334***	–1.803	–2.754	–0.470	–1.770
	(30.978)	(30.825)	(30.473)	(25.843)	(3.064)	(3.066)	(2.991)	(3.096)
Administration	–141.012***	–128.246**	–117.297**	–112.181**	1.256	–0.398	1.110	0.122
	(53.855)	(53.382)	(56.928)	(56.142)	(5.327)	(5.285)	(5.007)	(5.204)
Armed forces	69.074	73.641	59.046	–29.173	–0.527	–0.752	–0.086	–0.431
	(95.321)	(94.249)	(93.482)	(79.926)	(9.429)	(9.303)	(9.245)	(9.280)
ρ	–	0.076**	–	–0.643***	–	–0.094***	–	–0.061
		(0.038)		(0.055)		(0.034)		(0.050)
λ	–	–	0.224***	0.727***	–	–	–0.129***	–0.060
			(0.041)	(0.030)			(0.046)	(0.067)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
LL	–4,993.652	–4,991.275	–4,981.850	–4,944.904	–2,610.838	–2,606.795	–2,607.158	–2,606.468
AIC	10,023.300	10,022.550	10,003.700	9,931.809	5,257.676	5,253.591	5,254.315	5,254.935
Observation	1,030	1,030	1,030	1,030	1,030	1,030	1,030	1,030

(*Continued*)

TABLE 6
(*Continued*)

B: Measures of spatial effects

	Direct effect	Indirect effect	Total effect	Direct effect	Indirect effect	Total effect
ln(Population)	8.525	–3.770	4.755	2.963	–0.259	2.705
ln(Income)	48.328	–21.372	26.956	1.953	–0.171	1.783
Wholesale	–1,035.891	458.096	–577.795	–48.744	4.257	–44.487
Retail	–263.536	116.542	–146.994	–36.970	3.229	–33.742
Finance	374.466	–165.598	208.868	7.784	–0.680	7.104
Education	–162.955	72.063	–90.892	–2.760	0.241	–2.519
Administration	–122.413	54.134	–68.279	–0.398	0.035	–0.364
Armed forces	–31.834	14.078	–17.756	–0.754	0.066	–0.688

Note: We take the transformation of natural logarithm for two continuous variables (population and income). "FE" stands for the fixed effect; *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

Lastly, we observe in both cases that the direct and indirect effects of other industrial variables, in particular wholesale, retail, and education industries, move in the same direction of loss frequency and severity; that is, counties with higher proportion of such industries tend to be less affected by cyber loss events (with respect to both frequency and severity) themselves, whereas their neighboring counties tend to face high frequency. This result for loss frequency is in line with that of Table 4 for all of the counties.[37]

According to Shevchenko et al. (2021), risk types may lead to significant heterogeneity in the statistical features of cyber loss frequency and severity. For instance, malicious types including cyber extortion and distributed denial of service and privacy breach types can show high frequency, whereas high severity is observed mostly in malicious types. Adding further contribution to this literature, we reclassify cyber risks into three classes (malicious, unintended, and privacy breach) and examine whether heterogeneity in spatial effects exists among three classes and how heterogeneous the impacts of socioeconomic factors can be on cyber loss frequency (Table 7) and severity (Table 8).

It is observed that spatial effects are present for all risk types in the loss frequency model, with the SAC approach more competitive to address the impacts of socioeconomic factors than others, whereas the spatial effect is observed in the malicious and unintended risk type for the loss severity model. We find consistent results that socioeconomic factors representing the size of loss exposures (i.e., population and average income) overall show positive and significant impacts on both loss frequency and severity, except for the income variable of the unintended risk type in the loss frequency case of the optimal model (SAC). A possible implication for this result is that regions with large population may have higher likelihood of being exposed to unintended risk events caused by human errors; however, the economic development may not be associated with such likelihood.[38] We further identify that the coefficients of the financial industry variable are positive and significant in addressing loss frequency and severity of all three risk types at the county level. We can conclude that regardless of risk type, large and wealthy counties with more weight on the financial industry are more likely to be exposed to cyber risks (both in frequency and in severity).

5.3. Discussion

The common spatial features of cyber risk in terms of frequency and severity may have important implications for risk managers, policymakers, and risk management solution providers (e.g., cyber insurers). First, businesses in a region with a relatively large population or high income may need to be more attentive to cybersecurity enhancement and might need to pay higher premiums than other regions. Institutions and clients in such regions tend to be more exposed to cyber loss events, and businesses in those areas are more likely to become major targets of cyberattacks. Cybersecurity requirements for organizations may help control contagious cyber risks and reduce societal costs. Furthermore, regions characterized by a greater concentration of the financial industry may need higher cyber insurance premiums.

Second, as interconnectivity among industries continues to rise with the advent of digital transformation, mitigating cyber incidents through a limited focus on the cybersecurity levels of individual industries is challenging, and instead, an industry-wide elevation of cybersecurity standards is essential. This can be associated with our finding that regions with a high proportion of the financial industry tend to have high frequency and high severity of cyber risks. The features of the business model of the financial industry, for example, a "hub-and-spoke" type of the industrial infrastructure and intermediation-based business systems, may address such a finding that large loss amount can apply to interconnected regions with large-sized financial industries (Aldasoro et al. 2022).[39]

Additionally, our findings can contribute to complementing a typical approach of cyber model vendors or cyber insurers to building firmographic data to better understand risk profiles of potential victims exposed to cyber risks. Understanding risk profiles of corporate victims with firmographic data can help insurers improve their cyber risk assessment and thus pricing scheme. Typical firmographic data for cyber risks contain firm-specific information, for example, firm size (revenue or the

[37]To see whether lagged effects of key explanatory variables (i.e., population and average income) are valid, we also check cross correlations between their lagged variables and the dependent variable by testing statistical significance of 50 states and checking cross-correlation plots. We find that most cases across different lagged correlation show statistical insignificance. The results are available upon request.

[38]We check robustness of this result with alternative data of cyber risk (i.e., data breach risk) in Appendix C and find the consistent result that the income level in the model of the nonhacking risk type is not statistically significant in the loss frequency case of the optimal model (SAC).

[39]Aldasoro et al. (2022) further discuss the risk of a dense connectivity between financial companies by exemplifying heavy dependency on the cloud system. Potential resulting losses by such dependency in the financial system could be extreme, in particular, if the size of a shock to the network environment is large (Aldasoro et al. 2022).

TABLE 7
Results of Spatial Panel Analysis in Three Risk Types (Loss Frequency)

A: Comparison of regression results

	Malicious risk type				Unintended risk type				Privacy risk type			
	SLM	SAR	SEM	SAC	SLM	SAR	SEM	SAC	SLM	SAR	SEM	SAC
Const	–41.255***	–31.680***	–39.332***	–32.691***	–12.338***	–9.802***	–11.592***	–2.198***	–17.458***	–11.571***	–15.494***	–19.328***
	(1.840)	(1.786)	(1.785)	(1.497)	(0.815)	(0.802)	(0.803)	(0.678)	(1.807)	(1.777)	(1.774)	(1.664)
ln(Pop)	1.259***	1.017***	1.202***	1.057***	0.542***	0.469***	0.533***	0.188***	1.306***	1.133***	1.312***	1.407***
	(0.027)	(0.028)	(0.031)	(0.032)	(0.012)	(0.012)	(0.013)	(0.007)	(0.027)	(0.028)	(0.030)	(0.034)
ln(Inc)	3.204***	2.432***	3.089***	2.660***	0.807***	0.612***	0.750***	0.068	0.835***	0.371**	0.654***	1.027***
	(0.181)	(0.177)	(0.199)	(2.519)	(0.080)	(0.079)	(0.086)	(0.046)	(0.177)	(0.175)	(0.193)	(0.211)
WS	–17.258***	–17.096***	–19.437***	–16.607***	–4.087***	–4.573***	–5.319***	–2.015**	–5.029*	–6.280**	–8.745***	–11.669***
	(2.855)	(2.770)	(2.907)	(2.519)	(1.264)	(1.244)	(1.292)	(0.839)	(2.803)	(2.756)	(2.867)	(2.798)
Retail	–28.376***	–23.837***	–24.738***	–19.439***	–12.033***	–10.749***	–11.079***	–7.107***	–26.329***	–23.609***	–24.931***	–24.301***
	(1.484)	(1.442)	(1.480)	(1.241)	(0.657)	(0.648)	(0.662)	(0.463)	(1.457)	(1.435)	(1.465)	(1.391)
Finance	22.913***	20.961***	20.200***	16.940***	13.590***	12.605***	12.306***	10.026***	33.823***	31.774***	31.541***	28.963***
	(2.154)	(2.090)	(2.228)	(1.975)	(0.954)	(0.939)	(0.986)	(0.611)	(2.115)	(2.080)	(2.191)	(2.180)
Edu	–2.606***	–2.295***	–4.243***	–5.965***	–0.744**	–0.577*	–1.122***	0.301	–8.438***	–7.127***	–8.503***	–9.484***
	(0.747)	(0.726)	(0.782)	(0.706)	(0.331)	(0.326)	(0.345)	(0.205)	(0.733)	(0.723)	(0.767)	(0.775)
Admin	1.833	0.438	1.061	–0.132	0.656	0.290	0.460	–0.360	–0.455	–2.279**	–2.966**	–4.632***
	(1.136)	(1.102)	(1.227)	(1.167)	(0.503)	(0.495)	(0.536)	(0.293)	(1.115)	(1.097)	(1.197)	(1.265)
Armed	–14.775***	–11.642***	–12.676***	–10.247***	–5.449***	–4.658***	–5.035***	–2.255***	–6.392***	–4.442**	–4.955**	–4.622**
	(2.282)	(2.215)	(2.265)	(1.896)	(1.010)	(0.995)	(1.014)	(0.725)	(2.240)	(2.204)	(2.244)	(2.125)
ρ	–	0.267***	–	–0.861***	–	0.199***	–	0.777***	–	0.210***	–	–0.501***
		(0.008)		(0.012)		(0.008)		(0.006)		(0.008)		(0.018)
λ	–	–	0.278***	0.834***	–	–	0.199***	–0.885***	–	–	0.227***	0.626***
			(0.008)	(0.005)			(0.009)	(0.012)			(0.008)	(0.011)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
LL	–100,217.20	–99,486.48	–99,496.65	–98,525.23	–74,696.70	–74,332.90	–74,363.39	–73,324.33	–99,643.37	–99,251.58	–99,219.77	–99,059.27
AIC	200,470.30	199,013.00	199,033.30	197,092.50	149,429.40	148,705.80	148,766.80	146,690.70	199,322.70	198,543.20	198,479.50	198,160.50
Obs	31,320	31,320	31,320	31,320	31,320	31,320	31,320	31,320	31,320	31,320	31,320	31,320

(*Continued*)

TABLE 7
(*Continued*)

B: Measures of spatial effects

	Direct effect	Indirect effect	Total effect	Direct effect	Indirect effect	Total effect	Direct effect	Indirect effect	Total effect
ln(Pop)	1.189	–0.621	0.568	0.227	0.618	0.845	1.464	–0.527	0.937
ln(Inc)	2.991	–1.562	1.430	0.082	0.222	0.304	1.068	–0.384	0.684
WS	–18.674	9.749	–8.926	–2.429	–6.616	–9.045	–12.141	4.367	–7.773
Retail	–21.858	11.411	–10.447	–8.569	–23.336	–31.904	–25.283	9.095	–16.187
Finance	19.049	–9.944	9.105	12.089	32.923	45.012	30.132	–10.840	19.292
Edu	–6.707	3.501	–3.206	0.362	0.987	1.350	–9.867	3.550	–6.317
Admin	–0.148	0.077	–0.071	–0.434	–1.183	–1.617	–4.819	1.734	–3.086
Armed	–11.522	6.015	–5.507	–2.719	–7.404	–10.123	–4.809	1.730	–3.079

Note: We take the transformation of natural logarithm for two continuous variables (population and income). "Const," "Pop," "Inc," "WS," "Edu," "Admin," "Armed," "FE," and "Obs" stand for the constant term, population, income, wholesale, education, public administration, armed forces, fixed effect, and sample size, respectively; *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

TABLE 8
Results of Spatial Panel Analysis in Three Risk Types (Loss Severity)

A: Comparison of regression results

	Malicious risk type				Unintended risk type				Privacy risk type			
	SLM	SAR	SEM	SAC	SLM	SAR	SEM	SAC	SLM	SAR	SEM	SAC
Const	–29.408***	–26.706***	–28.317***	–28.980***	–15.515***	–13.855***	–14.963***	–15.316***	–12.847***	–9.836***	–12.031***	–11.972***
	(0.751)	(0.747)	(0.745)	(0.744)	(0.589)	(0.586)	(0.585)	(0.584)	(0.499)	(0.487)	(0.485)	(0.486)
ln(Pop)	1.069***	1.000***	1.085***	1.117***	0.671***	0.626***	0.680***	0.697***	0.575***	0.488***	0.581***	0.578***
	(0.011)	(0.012)	(0.012)	(0.015)	(0.009)	(0.009)	(0.009)	(0.012)	(0.007)	(0.008)	(0.008)	(0.010)
ln(Inc)	1.992***	1.770***	1.871***	1.917***	0.966***	0.833***	0.905***	0.929***	0.806***	0.569***	0.727***	0.723***
	(0.074)	(0.074)	(0.079)	(0.083)	(0.058)	(0.058)	(0.061)	(0.064)	(0.049)	(0.569)	(0.054)	(0.055)
WS	–4.754***	–4.724***	–5.101***	–5.191***	–0.923	–1.083	–1.442	–1.537	–0.813	–1.061	–1.526*	–1.516*
	(1.165)	(1.159)	(1.191)	(1.199)	(0.913)	(0.908)	(0.932)	(0.938)	(0.774)	(0.756)	(0.791)	(0.791)
Retail	–12.726***	–11.807***	–12.046***	–12.190***	–10.091***	–9.445***	–9.675***	–9.750***	–8.577***	–7.339***	–7.797***	–7.785***
	(0.606)	(0.604)	(0.612)	(0.617)	(0.475)	(0.473)	(0.479)	(0.484)	(0.402)	(0.394)	(0.403)	(0.405)
Finance	14.549***	14.050***	13.417***	13.218***	10.843***	10.374***	10.005***	9.889***	13.978***	12.698***	12.163***	12.187***
	(0.879)	(0.875)	(0.906)	(0.915)	(0.689)	(0.685)	(0.709)	(0.715)	(0.584)	(0.571)	(0.607)	(0.606)
Edu	0.827***	1.095***	0.705**	0.521	–0.016	0.185	–0.075	–0.173	–1.876***	–1.353***	–1.915***	–1.897***
	(0.305)	(0.304)	(0.316)	(0.323)	(0.239)	(0.238)	(0.247)	(0.252)	(0.202)	(0.198)	(0.213)	(0.215)
Admin	2.763***	2.259***	2.287***	2.320***	2.779***	2.383***	2.508***	2.547***	1.219***	0.521*	0.386	0.391
	(0.464)	(0.462)	(0.489)	(0.500)	(0.363)	(0.362)	(0.382)	(0.390)	(0.308)	(0.301)	(0.334)	(0.334)
Armed	–7.030***	–6.259***	–6.678***	–6.860***	–2.745***	–2.297***	–2.601***	–2.705***	–4.355***	–3.329***	–3.758***	–3.747***
	(0.931)	(0.928)	(0.938)	(0.942)	(0.730)	(0.727)	(0.735)	(0.738)	(0.618)	(0.605)	(0.616)	(0.617)
ρ	–	0.111***	–	–0.053***	–	0.117***	–	–0.044**	–	0.245***	–	0.009
		(0.008)		(0.015)		(0.008)		(0.020)		(0.008)		(0.018)
λ	–	–	0.155***	0.207***	–	–	0.142***	0.186***	–	–	0.283***	0.275***
			(0.009)	(0.016)			(0.009)	(0.020)			(0.008)	(0.018)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
LL	–72,147.06	–72,036.20	–71,988.81	–71,984.35	–64,511.13	–64,400.44	–64,378.40	–64,376.79	–59,320.47	–58,773.86	–58,724.77	–58,724.71
AIC	144,330.10	144,112.40	144,017.60	144,010.70	129,058.30	128,840.90	128,796.80	128,795.60	118,676.90	117,587.70	117,489.50	117,491.40
Obs	31,320	31,320	31,320	31,320	31,320	31,320	31,320	31,320	31,320	31,320	31,320	31,320

(*Continued*)

TABLE 8
(*Continued*)

B: Measures of spatial effects

	Direct effect	Indirect effect	Total effect	Direct effect	Indirect effect	Total effect	Direct effect	Indirect effect	Total effect
ln(Pop)	1.118	–0.057	1.061	0.698	–0.030	0.668	–	–	–
ln(Inc)	1.918	–0.098	1.820	0.929	–0.040	0.890	–	–	–
WS	–5.193	0.265	–4.928	–1.537	0.066	–1.472	–	–	–
Retail	–12.196	0.623	–11.573	–9.753	0.417	–9.336	–	–	–
Finance	13.225	–0.676	12.549	9.892	–0.423	9.469	–	–	–
Edu	0.522	–0.027	0.495	–0.173	0.007	–0.166	–	–	–
Admin	2.321	–0.119	2.203	2.548	–0.109	2.439	–	–	–
Armed	–6.864	0.351	–6.513	–2.706	0.116	–2.591	–	–	–

Note: We take the transformation of natural logarithm for two continuous variables (population and income). "Const," "Pop," "Inc," "WS," "Edu," "Admin," "Armed," "FE," and "Obs" stand for the constant term, population, income, wholesale, education, public administration, armed forces, fixed effect, and sample size, respectively; *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively. Panel B of privacy risk type models does not provide the results because the SEM model presents the lowest AIC; hence no evidence on spatial effects is observed.

number of employees), industry type, geographic location, and cyber risk governance, as relatively more important factors than others (Guy Carpenter 2022). Our findings of spatial dependency and socioeconomic factors with spatial effects may argue that there is a need to consider possible interaction between firmographic factors, for example, industry and geographical location, and this interaction can help vendors and insurers make geographical landscape of cyber losses more concrete.

In this sense, our results imply for insurance practitioners that cyber insurance pricing needs to be differentiated by socioeconomic factors such as size of exposures and industry types. Eling et al. (2022) already illustrate that industrial factors can be significant in premium calculation by showing the difference in cyber insurance premiums between financial and medical firms. This study adds more insight for cyber insurers by highlighting regional risk levels relying on population and wealth levels. Carriers in the current cyber insurance markets tend to consider firm size, industry, and security level as primary risk factors (see, e.g., Zeller and Scherer 2022; Awiszus et al. 2023). However, spatial clusters of our study may signal higher exposures to systematic cyber risks, implying that entities in the hot clusters may be more exposed to systematic cyber losses (Awiszus et al. 2023).

In association with the preceding implications, this study is among the first to provide an academic analysis to show spatial loss clusters that consider regional vulnerability to cyber risks and their socioeconomic drivers. The literature in both industry and academia tends to focus on (1) geographical contiguity of cyber loss frequency between regions and only the patterns by economic development and population (see, e.g., Khey and Sainato 2013; Chen et al. 2021) or (2) network dependency because cyber risk events are rooted across borders via the network structure (see, e.g., Hua and Xu 2021; Hillairet et al. 2022; World Economic Forum 2022, Chapter 3). We contribute additional insights to this literature by emphasizing that the examination of spatial patterns, particularly at the county level, can help address the micro landscape of cyber losses, and cyber loss exposures can be socioeconomically affected.

6. CONCLUSION

Combining cyber loss information and county-level socioeconomic information, we investigate spatial dependency in cyber loss frequency and severity of U.S. states and counties. Global and local Moran's I statistics to measure autocorrelation reveal statistical significance of county-level spatial dependency in cyber loss frequency and severity. However, we find no statistical evidence on state-level spatial dependency, implying that cyber risk events tend to occur within targeted states but are less likely to happen between adjacent states. We then use three spatial econometric approaches (SAR, SEM, and SAC) to investigate the spatial impacts of socioeconomic factors on cyber loss frequency and severity at the county level, with factors that are average income, population as a proxy for Internet usage, and the size of cyber loss exposures and industry identifiers. We consistently find that population size and average income are positive and significant in explaining both frequency and severity, implying that large and wealthy counties are more exposed to cyber risks.

We further analyze the cases that are in regions relatively exposed (hot clusters) to cyber risks based on local Moran's I analysis and three risk classes. We determine that the financial industry of the Northeast Coast area tends to experience more cyber risks than the state of California. When it comes to the heterogeneous risk type analysis, we find a spatial effect for all risk types in both loss frequency and severity. Our findings addressing common features of cyber loss frequency and severity in spatial approaches can help policymakers, risk managers, and risk management providers to better understand spatial patterns of cyber risk events and key socioeconomic factors that need to be considered.

Although this study can provide useful information on how the frequency and severity of cyber risk events are related geographically and what factors may drive the geographical patterns, there is a potential avenue to improve our understanding of what spatial features cyber risk events have. First, we still need to figure out how close different events are to each other within the network. To achieve this, there is a need to develop a quantitative metric to measure network distance between cyber risk events, which can overcome the limitation of spatial dependency measuring physical distance. In addition, we may need to consider further socioeconomic information, for example, the intensity of Internet usage, regional investments on cyber security systems, or population distribution over ages.

Our spatial models investigating the impacts of socioeconomic factors of population as a proxy for Internet use or income levels are required to make some types of assumptions when one interprets the findings or spatial patterns of dependence such that the companies of importance with regard to cyberattacks and the relevance of their data are related to counties with high population density or high income level. For example, this may fail to account for the fact that high-income workers might live in different locations than their employer (where the cyber event would typically take place); hence, this might potentially reveal false assumptions on the spatial correlation analysis.

It would also be interesting to more deeply study individual victims of cyber loss events and their spatial features, if such data are available. Another limitation of our study is the consideration of yearly data in the spatial regressions, which might

open avenues for future research once respective information becomes available. Cyber insurance markets and policymakers can take advantage of such additional information to expand their understanding of cyber risk drivers and improve decisions on cyber security.

DISCLOSURE STATEMENT

There are no financial or nonfinancial competing interests to report.

FUNDING

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023S1A5A8076897).

REFERENCES

Aldasoro, I., L. Gambacorta, P. Giudici, and T. Leach. 2022. The drivers of cyber risk. *Journal of Financial Stability* 60:100989.

Amin, R. W., H. E. Sevil, S. Kocak, G. Francia, and P. Hoover. 2021. The spatial analysis of the malicious Uniform resource locators (URLs): 2016 Dataset case study. *Information* 12 (1):2. doi:10.3390/info12010002

Anselin, L. 1995. Local indicators of spatial association-LISA. *Geographical Analysis* 27 (2):93–115. doi:10.1111/j.1538-4632.1995.tb00338.x

Anselin, L., and S. J. Rey. 2014. *Modern spatial econometrics in practice: A guide to GeoDa, GeoDaSpace and PySAL*. GeoDa Press LLC.

Anselin, L., I. Syabri, and Y. Kho. 2006. GeoDa: An introduction to spatial data analysis. *Geographical Analysis* 38 (1):5–22. doi:10.1111/j.0016-7363.2005.00671.x

Awiszus, K., T. Knispel, I. Penner, G. Svindland, A. Voß, and S. Weber. 2023. Modeling and pricing cyber insurance: Idiosyncratic, systematic and systemic risks. *European Actuarial Journal* 13:1–53. doi:10.1007/s13385-023-00341-9

Baldwin, A., I. Gheyas, C. Ioannidis, D. Pym, and J. Williams. 2017. Contagion in cyber security attacks. *Journal of the Operational Research Society* 68 (7):780–91. doi:10.1057/jors.2016.37

Biener, C., M. Eling, and J. H. Wirfs. 2015. Insurability of cyber risk: An empirical analysis. *Geneva Papers on Risk and Insurance—Issues and Practice* 40 (1):131–58. doi:10.1057/gpp.2014.19

Carr, J. 2012. *Inside cyber warfare: Mapping the cyber underworld*. Sebastopol: O'Reilly Media, Inc.

Castells, M. 2004. *The rise of the network society*, 2nd ed. Oxford: Blackwell.

Chen, S., C. Gao, D. Jiang, M. Hao, F. Ding, T. Ma, S. Zhang, and S. Li. 2021. The spatiotemporal pattern and driving factors of cyber fraud crime in China. *ISPRS International Journal of Geo-Information* 10 (12):802. doi:10.3390/ijgi10120802

Cotteleer, G., and G. C. van Kooten. 2012. Expert opinion versus actual transaction evidence in the valuation of non-market amenities. *Economic Modelling* 29 (1):32–40. doi:10.1016/j.econmod.2011.02.041

Darmofal, D. 2015. *Spatial analysis for the social sciences*. New York: Cambridge University Press.

Edwards, B., S. Hofmeyr, and S. Forrest. 2016. Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity* 2 (1):3–14. doi:10.1093/cybsec/tyw003

Eisenbach, T. M., A. Kovner, and M. J. Lee. 2022. Cyber risk and the US financial system: A pre-mortem analysis. *Journal of Financial Economics* 145 (3):802–26. doi:10.1016/j.jfineco.2021.10.007

Elhorst, J. P. 2013. Spatial panel models. In *Handbook of regional science*, ed. M. M. Fischer and P. Nijkamp, vol. 3, 1637–52. Heidelberg: Springer.

Elhorst, J. P. 2014. *Spatial econometrics: From cross-sectional data to spatial panels*. Heidelberg: Springer.

Eling, M., and N. Loperfido. 2017. Data breaches: Goodness of fit, pricing, and risk measurement. *Insurance: Mathematics and Economics* 75:126–36. doi:10.1016/j.insmatheco.2017.05.008

Eling, M., and K. Jung. 2018. Copula approaches for modeling cross-sectional dependence of data breach losses. *Insurance: Mathematics and Economics* 82:167–80. doi:10.1016/j.insmatheco.2018.07.003

Eling, M., and K. Jung. 2022. Heterogeneity in cyber loss severity and its impact on cyber risk measurement. *Risk Management* 24 (4):273–97. doi:10.1057/s41283-022-00095-w

Eling, M., K. Jung, and J. Shim. 2022. Unraveling heterogeneity in cyber risks using quantile regressions. *Insurance: Mathematics and Economics* 104:222–242. doi:10.1016/j.insmatheco.2022.03.001

Eling, M., and J. H. Wirfs. 2019. What are the actual costs of cyber risk events? *European Journal of Operational Research* 272 (3):1109–19. doi:10.1016/j.ejor.2018.07.021

Felson, M. 2008. Routine activity approach. In *Environmentalcriminology and crime analysis*, ed. R. Worley and L. Mazerolle, 70–77. Cullompton, UK: Willan.

Fischer, M. M., and J. Wang. 2011. *Spatial data analysis: Models, methods and techniques*. New York: Springer Science & Business Media.

Golgher, A. B., and P. R. Voss. 2016. How to interpret the coefficients of spatial models: Spillovers, direct and indirect effects. *Spatial Demography* 4 (3):175–205. doi:10.1007/s40980-015-0016-y

Gorman, S. P., and E. J. Malecki. 2000. The networks of the Internet: An analysis of provider networks in the USA. *Telecommunications Policy* 24 (2):113–34. doi:10.1016/S0308-5961(00)00005-7

Guy Carpenter. 2022. *Decoding cyber risk*. CAS Casualty Loss Reserve Seminar. New York: Guy Carpenter.

Hillairet, C., O. Lopez, L. d'Oultremont, and B. Spoorenberg. 2022. Cyber-contagion model with network structure applied to insurance. *Insurance: Mathematics and Economics* 107:88–101. doi:10.1016/j.insmatheco.2022.08.002

Hua, L., and M. Xu. 2021. Pricing cyber insurance for a large-scale network. *Variance* 14 (2):1–19.
Jung, K. 2021. Extreme data breach losses: An alternative approach to estimating probable maximum loss for data breach risk. *North American Actuarial Journal* 25 (4):580–603. doi:10.1080/10920277.2021.1919145
Kelejian, H. H., and I. R. Prucha. 1998. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics* 17 (1):99–121.
Khey, D. N., and V. A. Sainato. 2013. Examining the correlates and spatial distribution of organizational data breaches in the United States. *Security Journal* 26 (4):367–82. doi:10.1057/sj.2013.24
Kigerl, A. 2012. Routine activity theory and the determinants of high cybercrime countries. *Social Science Computer Review* 30 (4), 470–86. doi:10.1177/0894439311422689
LeSage, J. P., and R. K. Pace. 2009. *Introduction to spatial econometrics*. Boca Raton, FL: Taylor & Francis.
Li, H., C. A. Calder, and N. Cressie. 2007. Beyond Moran's *I*: Testing for spatial dependence based on the spatial autoregressive model. *Geographical Analysis* 39 (4):357–75. doi:10.1111/j.1538-4632.2007.00708.x
Ma, B., T. Chu, and Z. Jin. 2022. Frequency and severity estimation of cyber attacks using spatial clustering analysis. *Insurance: Mathematics and Economics* 106:33–45. doi:10.1016/j.insmatheco.2022.04.013
Malavasi, M., G. W. Peters, P. V. Shevchenko, S. Trück, J. Jang, and G. Sofronov. 2022. Cyber risk frequency, severity and insurance viability. *Insurance: Mathematics and Economics* 106:90–114. doi:10.1016/j.insmatheco.2022.05.003
Malecki, E. J. 2002. Hard and soft networks for urban competitiveness. *Urban Studies* 39 (5–6):929–45. doi:10.1080/00420980220128381
Munro, A. 2016, April 21. *Is Wall Street cyber secure?* Global Risk Insights. https://globalriskinsights.com/2016/04/wall-street-cyber-security.
Namestnikov, Y. 2012, September 11. *The geography of cybercrime: Western Europe and North America*. Securist. https://securelist.com/the-geography-of-cybercrime-western-europe-and-north-america/36671.
Okunlola, O. A., M. Alobid, O. E. Olubusoye, K. Ayinde, A. F. Lukman, and I. Szűcs. 2021. Spatial regression and geostatistics discourse with empirical application to precipitation data in Nigeria. *Scientific Reports*, 11(1), 1–14. doi:10.1038/s41598-021-96124-x
Palsson, K., S. Gudmundsson, and S. Shetty. 2020. Analysis of the impact of cyber events for cyber insurance. *Geneva Papers on Risk and Insurance—Issues and Practice* 45 (4):564–79. doi:10.1057/s41288-020-00171-w
Park, J., D. Cho, J. K. Lee, and B. Lee. 2019. The economics of cybercrime: The role of broadband and socioeconomic status. *ACM Transactions on Management Information Systems (TMIS)* 10 (4):1–23. doi:10.1145/3351159
Ponemon Institute. 2022. *Cost of a data breach report 2022*. Ponemon Institute LLC. https://www.ibm.com/security/data-breach.
Priemus, H. 2007. The network approach: Dutch spatial planning between substratum and infrastructure networks. *European Planning Studies* 15 (5):667–86. doi:10.1080/09654310701213962
Ramajo, J., M. A. Marquez, G. J. Hewings, and M. M. Salinas. 2008. Spatial heterogeneity and interregional spillovers in the European Union: Do cohesion policies encourage convergence across regions? *European Economic Review* 52 (3):551–67. doi:10.1016/j.euroecorev.2007.05.006
Reyns, B. W. 2017. Routine activity theory and cybercrime: A theoretical appraisal and literature review. In *Technocrime and criminological theory*, ed. K. F. Steinmetz and M. R. Nobles, 35–54. New York: Routledge.
Romanosky, S. 2016. Examining the costs and causes of cyber incidents. *Journal of Cybersecurity* 2 (2):121–35. doi:10.1093/cybsec/tyw001
Rutherford, J., A. Gillespie, and R. Richardson. 2004. The territoriality of pan-European telecommunications backbone networks. *Journal of Urban Technology* 11 (3):1–34. doi:10.1080/10630730500064166
Rusche, K. 2010. Quality of life in the regions: An exploratory spatial data analysis for West German labor markets. *Review of Regional Research* 30 (1):1–22. doi:10.1007/s10037-009-0042-6
Sassen, S. 2001. *The global city: New York, London, Tokyo*, 2nd ed. Princeton, NJ: Princeton University Press.
Schmidtke, H. R. 2018. Is the Internet spatial? *Journal of Reliable Intelligent Environments* 4 (3):123–29. doi:10.1007/s40860-018-0064-3
Sen, R., and S. Borle. 2015. Estimating the contextual risk of data breach: An empirical approach. *Journal of Management Information Systems* 32 (2):314–41. doi:10.1080/07421222.2015.1063315
Shevchenko, P. V., J. Jang, M. Malavasi, G. Peters, G. Sofronov, and S. Trück. 2021. Quantification of cyber risk–risk categories and business sectors. SSRN 3858608.
Shirabe, T. 2005. A model of contiguity for spatial unit allocation. *Geographical Analysis* 37 (1):2–16. doi:10.1111/j.1538-4632.2005.00605.x
Tranos, E. 2013. *The geography of the Internet: Cities, regions and Internet infrastructure in Europe*. Cheltenham, UK: Edward Elgar.
Verizon. 2019. *2019 Data breach investigations report*. New York: Verizon. https://www.verizon.com/business/resources/reports/2019-data-breach-investigations-report.pdf.
Vojinovic, I. 2022. 10 of the biggest data breaches in history. Dataprot. https://dataprot.net/articles/biggest-data-breaches.
Voss, P. R., D. D. Long, R. B. Hammer, and S. Friedman. 2006. County child poverty rates in the US: a spatial regression approach. *Population Research and Policy Review* 25 (4):369–91. doi:10.1007/s11113-006-9007-4
Wheatley, S., T. Maillart, and D. Sornette. 2016. The extreme risk of personal data breaches and the erosion of privacy. *European Physical Journal B* 89 (1):1–12.
Windrem, R. 2015, July 31. Exclusive: Secret NSA map shows China cyber attacks on US target. *NBC News*. https://www.nbcnews.com/news/us-news/exclusive-secret-nsa-map-shows-china-cyber-attacks-us-targets-n401211.
World Economic Forum. 2022. *The global risks report 2022*. Geneva: World Economic Forum.
Zeller, G., and M. Scherer. 2022. A comprehensive model for cyber risk based on marked point processes and its application to insurance. *European Actuarial Journal* 12 (1):33–85. doi:10.1007/s13385-021-00290-1

APPENDIX A. RISK TYPE CLASSIFICATION

In Section 5.2, we explore the relationships between socioeconomic factors and cyber losses by different risk types. To do so, we first define three risk types that have occurred more frequently than others. Figure A1 shows the number of cyber risk events by 15 risk types that the dataset provides. We find that privacy, malicious, and unintended risk types cause loss events most frequently among others. Following Malavasi et al. (2022) on risk classification for the types of cyber threats, we denote the three risk classes as malicious, unintentional breach, and cyber risk events associated with privacy information. We classify all risk types of Figure A1 into these three classes (see Table A1), except for the undetermined case, which may lead to potential bias due to lack of information in risk characteristics (Eling and Jung 2018).

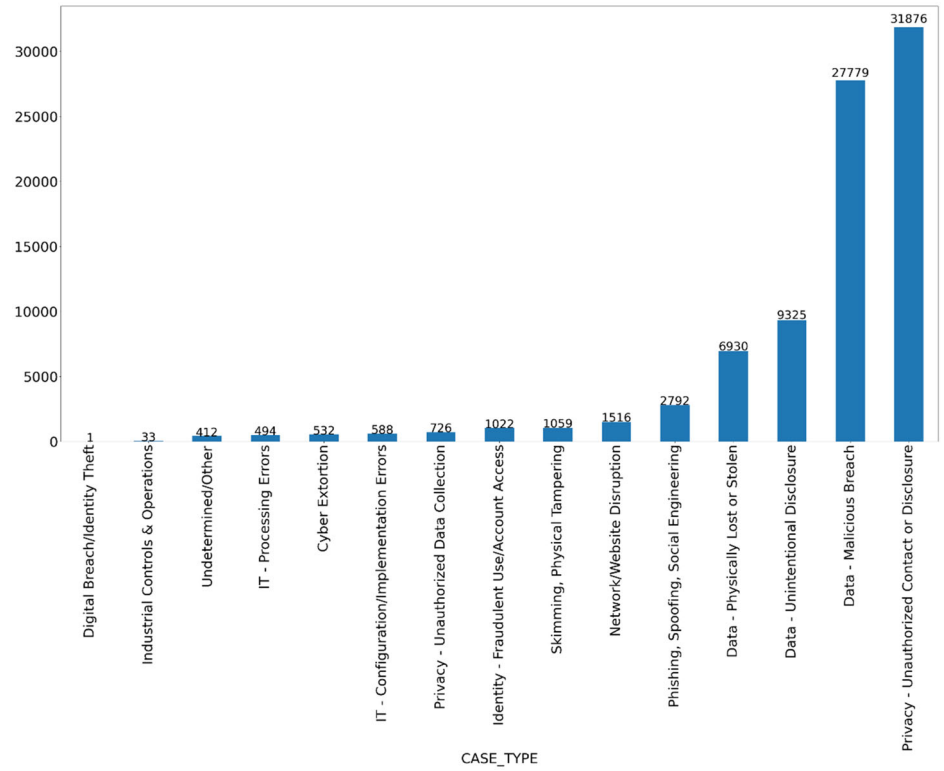

FIGURE A1. The number of loss events by risk type for the period from 2005 to 2020.

TABLE A1
Risk Type Classification

Risk category	Types of cyber risk events
Malicious	Cyber extortion
	Data—malicious breach
	Industrial controls and operations
	Network/website disruption
	Phishing, spoofing, social engineering
	Skimming, physical tampering
Unintended	Data—physically lost or stolen
	Data—unintentional disclosure
	IT—configuration/implementation errors
	IT—processing errors
Privacy	Privacy—unauthorized contact or disclosure
	Privacy—unauthorized data collection
	Identity—fraudulent use/account access
	Digital breach/identity theft

APPENDIX B. FORMULATION OF DIRECT AND INDIRECT EFFECTS

In this section, we briefly discuss how direct and indirect effects of spatial econometric models can be formulated and how they can be interpreted. First, regarding the SEM, we interpret the coefficients of the variables as aligned with that of the SLM. This is because the spatial coefficient of the SEM is linked with the error term, which follows the normal distribution with zero mean and variance σ_ϵ^2 as the SLM. Consequently, the expected value and partial derivatives for the dependent variable in the SEM mirror those of the SLM (Golgher and Voss 2016). However, the SAR and SAC present spatial coefficients associated with a dependent variable, where direct and indirect effects of an explanatory variable of a specific region on dependent variables of neighboring regions can be observed (Golgher and Voss 2016).

To understand the direct and indirect effects, we first define the matrix form of the SAR model:

$$y = \rho \boldsymbol{W}\boldsymbol{y} + \beta_0 \boldsymbol{i_n} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{B1}$$

where y is a n-dimensional vector of the dependent variable, where n indicates the number of regions in the model; ρ is a spatial coefficient and $\boldsymbol{W}$ is a $n \times n$ spatial weight matrix; i_n means an n-dimensional vector of ones and β_0 is an intercept of a spatial regression; and $\boldsymbol{X}$ is a $n \times p$ matrix, where p is the number of explanatory variables, $\boldsymbol{\beta}$ is a p-dimensional vector of the explanatory variables, and $\boldsymbol{\epsilon}$ presents a white noise error, which follows a normal distribution with zero mean and variance σ_ϵ^2.
Equation (B1) can be reformulated as follows:

$$y = (\boldsymbol{I} - \rho \boldsymbol{W})^{-1}(\beta_0 \boldsymbol{i_n} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}). \tag{B2}$$

We can derive the expected value of a dependent variable and partial derivatives for a kth explanatory variable from Equation (B2) in the following:

$$E(y|\boldsymbol{X}) = (\boldsymbol{I} - \rho \boldsymbol{W})^{-1} E[(\beta_0 \boldsymbol{i_n} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] = (\boldsymbol{I} - \rho \boldsymbol{W})^{-1}(\beta_0 \boldsymbol{i_n} + \boldsymbol{X}\boldsymbol{\beta}) \tag{B3}$$

$$\begin{pmatrix} \frac{\partial y_1}{\partial x_{1k}} & \cdots & \frac{\partial y_1}{\partial x_{nk}} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_{1k}} & \cdots & \frac{\partial y_n}{\partial x_{nk}} \end{pmatrix} = (\boldsymbol{I} - \rho \boldsymbol{W})^{-1} \begin{pmatrix} \beta_k & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \beta_k \end{pmatrix} = \beta_k (\boldsymbol{I} - \rho \boldsymbol{W})^{-1}$$

$$= \beta_k \begin{pmatrix} m_{11} & \cdots & m_{1n} \\ \vdots & \ddots & \vdots \\ m_{n1} & \cdots & m_{nn} \end{pmatrix},$$

where m_{ij} is an element of the $n \times n$ matrix, $(I - \rho W)^{-1}$.

Equation (B3) can lead us to observe spatial spillover effects in the SAR or SAC by obtaining non-zero values of off-diagonal elements of $(I - \rho W)^{-1}$. It is important to note that the elements of $(I - \rho W)^{-1}$ and the spatial weight matrix, W, are unit specific, implying that the direct and indirect effects vary across different units (Elhorst 2014).[40] LeSage and Pace (2009) propose a measure to quantify the direct and indirect effects across regions in a simple manner as follows:

$$\text{Total effect} = \frac{1}{n} \boldsymbol{i_n^T} \left(\beta_k (\boldsymbol{I} - \rho \boldsymbol{W})^{-1} \right) \boldsymbol{i_n} \tag{B4}$$

$$\text{Direct effect} = \frac{1}{n} \text{tr} \left(\beta_k (\boldsymbol{I} - \rho \boldsymbol{W})^{-1} \right) \tag{B5}$$

$$\text{Indirect effect} = \text{Total effect} - \text{Direct effect} \tag{B6}$$

That is, the authors propose an index for individual direct effects that can be calculated as the average of the diagonal elements of the partial derivatives, thus the average can influence over each region as in Equation (B5). They also suggest a measure for indirect effects that can be calculated as the average of the row sums (or columns sums) of the off-diagonal

[40]For example, in the case of N regions and k independent variables, there are k distinct $N \times N$ matrices representing direct and indirect effects.

elements of the partial derivatives, thus the average spillover effects can be derived in each region. One can define the total effect as the sum of the direct effect and indirect effect.

APPENDIX C. ROBUSTNESS CHECK WITH A DATA BREACH RISK DATASET

C1. Data Description

We investigate whether a similar pattern in spatial dependency of cyber risks and the impact of socioeconomic factors can be observed in another dataset. We use a public dataset of 9015 data breach events provided by the PRC. The PRC is a U.S. nonprofit organization and provided information on data breach events from 2005 to 2019, including 12 variables, such as victim organizations, industry, breach type, the number of breached records, and geographical locations (see panel A of Table C1). This dataset has been widely used in the literature (see, e.g., Edwards et al. 2016; Eling and Loperfido 2017; Eling and Jung 2018; Kamiya et al. 2021; Florackis et al. 2023) and is one of the largest public datasets for data breach risk.[41]

We use the breach type as a categorical variable (see panel B of Table C1), which helps us understand what type of risk can be more threatening and whether risk clusters may exist across the data breach risk landscape. The industry variable shows eight industrial types that constitute major areas of businesses and public sectors (see panel C of Table C1). The literature documents that the industry factor is significant in explaining vulnerability to cyber risks since homogeneity in business models and more frequent interactions are present in the identical industries (see, e.g., Wheatley et al. 2016; Zeller and Scherer 2022).

The total records variable is a continuous variable representing the number of breached records by a data breach event. This variable describes how many personal records are breached and multiple records may belong to the same victims. In this study, we use this variable to examine spatial features of data breach events by modeling data breach events for a certain time frame. Lastly, the variable for geographical locations is the key to our spatial models, the variable that presents latitude and longitude of an event. This variable can facilitate our model to investigate county-level features, which distinguishes our contribution from the relevant literature on the spatial modeling of cyber risks. This specific information with latitude and longitude is offered only by the PRC dataset.

TABLE C1
Variable Description and Risk/Industry Types

A: Variable description

Class	Column	Summary
Temporal	Year of breach	Year of data breach occurrence
	Date made public	Date of data breach announcement
Spatial	State	State of the location where data breach happened
	City	City of the location where data breach happened
	Latitude	Latitude of the location where data breach happened
	Longitude	Longitude of the location where data breach happened
Descriptive	Breach type	Category of data breach
	Industry type	Category of affected organization
	Total records	Number of breached records
	Company	Name of victim organization
	Description of incident	Details on the event
	Information source	Source that provides the event description

Continued

[41]One advantage of this dataset is that it is more homogeneous compared with other datasets used in the literature (e.g., SAS or Advisen), because it focuses on one type of risk (data breaches) and one country (United States). It is also less prone to biases than those other datasets, given that data breach notification has become mandatory in all U.S. states over the last years; see NCSL (2022). Other datasets are plagued by reporting biases, because many firms are reluctant to report cyber events. We note though that a data breach notification law has been in effect, not at the same time for all states, but in different time frames from 2002 to recently.

TABLE C1
(*Continued*)

B: Types of data breach

Type	Summary	Type	Summary
CARD	Debit/credit card fraud	PORT	Loss of portable device(s)
HACK	Hacking by outside/malware	STAT	Stationary computer loss
INSD	Insider of the organization	DISC	Unintended disclosure of data
PHYS	Physical damage/loss	UNKN	Unknown

C: Types of organization

Type	Summary	Type	Summary
BSF	Financial services	GOV	Government, utility
BSR	Retailers	MED	Healthcare/medical service provider
BSO	Other businesses	NGO	Nonprofit organization
EDU	Educational institution	UNKN	Unknown

We select observations of interest from the PRC dataset for the spatial dependency analysis of the study. First, we eliminate the unknown type of both risk type and industry variables since the type does not provide information on risk type or industry that are critical in our study. We further eliminate observations without geographical information (latitude and longitude), the data points that are mostly for the year 2019. Consistent with the sample selection in Section 4, we also consider the states with FIPS codes to combine the cyber loss data and the socioeconomic data. This process of elimination eventually leads us to have 5735 observations for the period from 2005 to 2018.

C2. Robustness Check for Spatial Dependency

We check the robustness of spatiality in relation to cyber risk events. Table C2 shows global Moran's *I* statistics representing testing results of state- and county-level spatial dependencies, respectively, for the three periods as done in Table 3. We clearly identify that there is no statistical evidence on spatial dependency across states over the period, but significant evidence on the dependency across counties at the 1% significance level. This result can also be addressed in Figure C1 that the fitted line of the left plot representing the state-level is nearly flat, whereas that of the right plot representing the county-level is linearly on the rise.

TABLE C2
State- and County-level Global Moran's I Statistics

	Loss frequency		Loss severity	
Year	State level	County level	State level	County level
Before 2014	–0.029	0.173***	–0.024	0.060***
After 2014	–0.059	0.139***	0.148**	–0.001
Entire period	–0.045	0.169***	0.135**	0.011**

Note: The figures indicate Moran's *I* statistics at the state and county level; *, **, and ***, indicate significance level of 10%, 5%, and 1%. We use the pseudo *p* value with GeoDa to test the statistics.

In addition, this county-level result illustrates that all significant spatial dependencies are positive. This result demonstrates that more exposed counties (i.e., counties with a large number of data breach events) and less exposed counties (i.e., counties with a small number of data breach events) are clustered. It can be implied that data breach events are geographically clustered

in particular locations, where certain socioeconomic environments may motivate outside attackers or there could be more risk exposures (e.g., large number of employees or more interconnected industries) potentially causing data disclosure risks. However, although statistical significance of loss severity at the county level is present, its slope in Figure C1 is overall flat. Thus, unlike the main body of the article, this section focuses on the loss frequency for spatial panel models to check robustness of main results. Following the preceding approach, we find consistent results with those of Section 5.1 that West and Northeast Coast regions (in particular, counties in California and New York) are more exposed to data breach events. Those counties are mostly categorized as high–high clusters, where highly exposed clusters to data breaches are present.

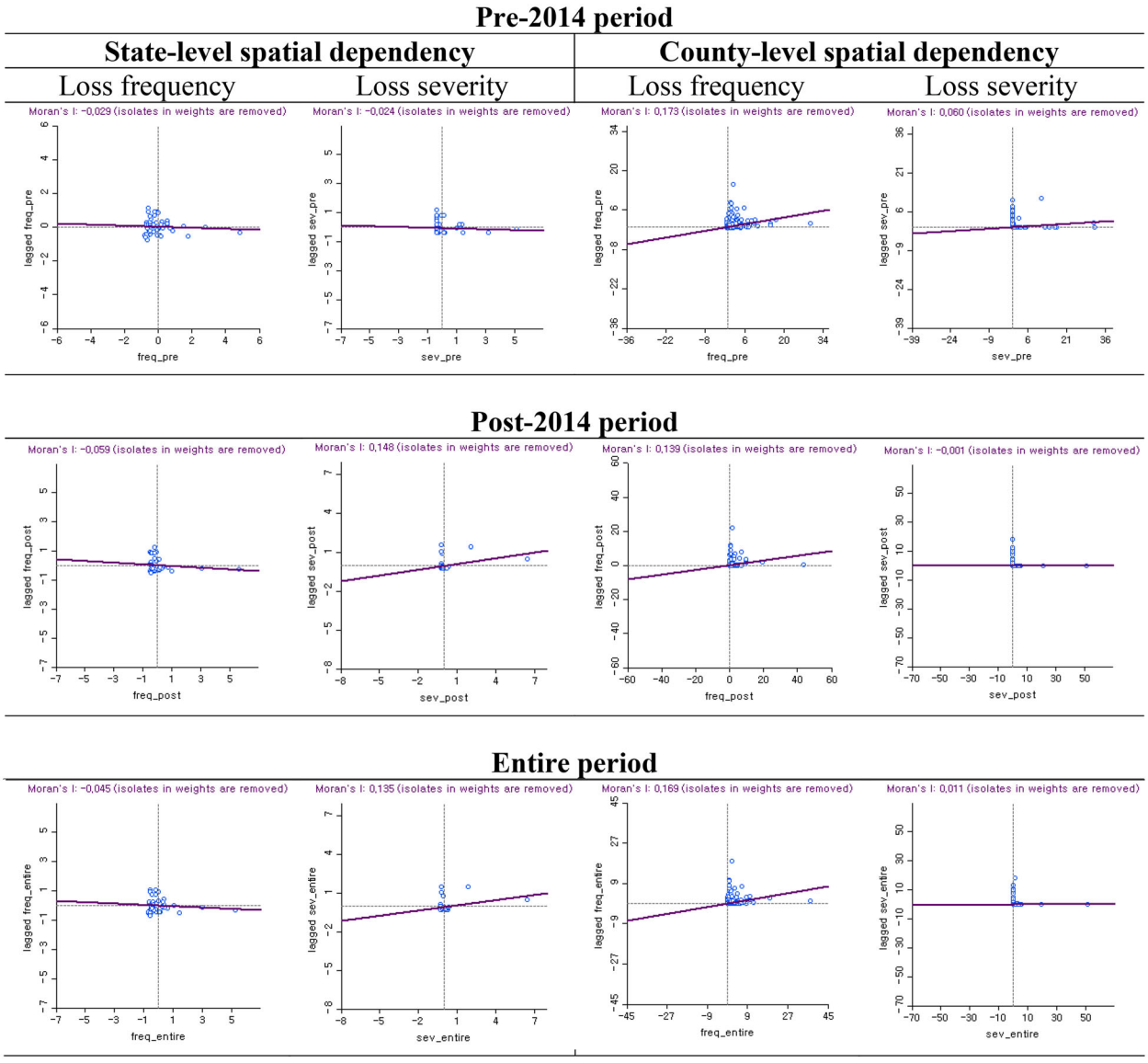

FIGURE C1. Global Moran's I with state-level data and county-level data

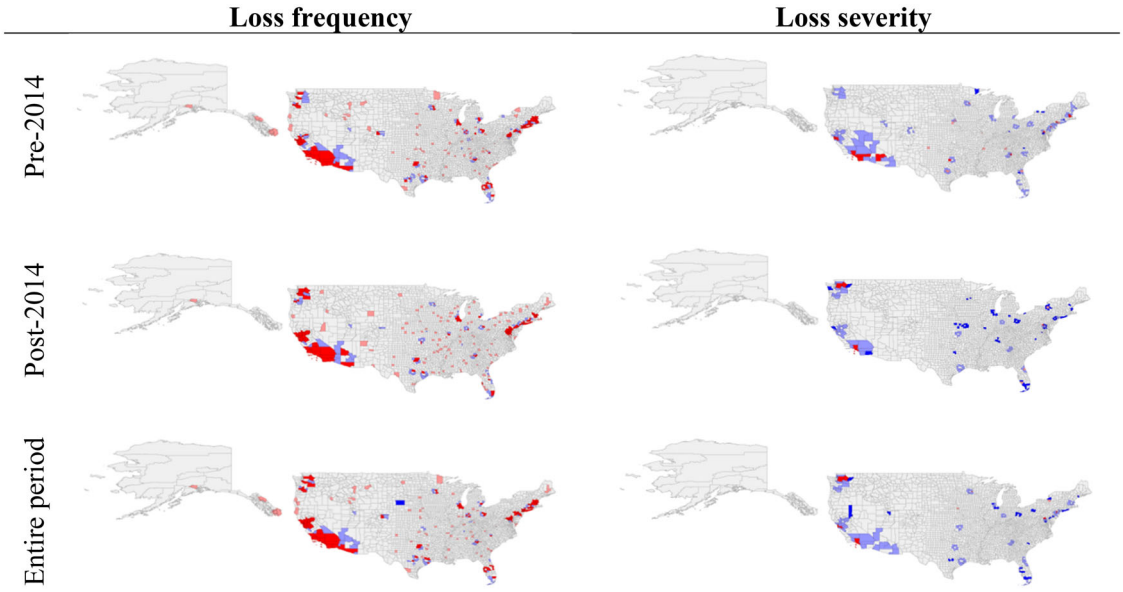

FIGURE C2. Results of Spatial Hot and Cold Spots with Regard to Data Breach Events at the County Level. *Note:* The graphs of the left (right) panel illustrate spatial dependency based on data breach frequency (severity). The categories in the legend indicate the following: high–high (red), the centered region and its adjacent counties all exposed to large number of data breach events; high–low (light red), the centered region highly exposed to data breach events, but its neighbors exposed to small number of events; low–high (light blue), the centered region less exposed to data breach events, but its neighbors exposed to large number of events; low–low (blue), the centered region and its adjacent counties all less exposed to data breach events; not significant, the regions that turn out to be insignificant in the statistical test.

C3. Robustness Check for Spatial Panel Models

Combining the PRC data and county-level socioeconomic data, we consider the period between 2011 and 2017 in 3133 counties of 50 states for this study.[42] Table C3 presents the results of four models (SLM and three spatial models) for all counties of the dataset. We find a consistent result for the model selection that the SAC model better explains spatial relationships between data breach frequency and socioeconomic factors than other models under the decision criteria of largest loglikelihood and smallest AIC. Our results also provide supporting evidence on significant spatial impacts of socioeconomic factors, particularly positive and significant impacts of cyber loss exposure proxies (population and average income).

TABLE C3
Results of Spatial Panel Analysis on All Counties (Loss Frequency)

A: Comparison of regression results

	Dependent variable: Data breach event frequency at the county level			
	SLM	SAR	SEM	SAC
Constant	–5.587***	–5.242***	–5.679***	–1.125***
	(0.504)	(0.502)	(0.503)	(0.422)
ln(Population)	0.223***	0.209***	0.219***	0.083***
	(0.008)	(0.008)	(0.008)	(0.004)
ln(Income)	0.404***	0.379***	0.416***	0.056**
	(0.050)	(0.050)	(0.051)	(0.028)

(*Continued*)

[42]For the PRC data, the sample size for 2018 and 2019 is limited to model the county-level spatial analysis; hence the inclusion of the two year sample may cause a bias in the model results, which leads us not to consider in this analysis. Unlike the major analysis with the Advisen dataset in Section 5.2, the PRC analysis includes the county of *Rio Arriba* since it has complete information on socioeconomic factors up to 2017.

TABLE C3
(*Continued*)

A: Comparison of regression results

	Dependent variable: Data breach event frequency at the county level			
	SLM	SAR	SEM	SAC
Wholesale	–2.502***	–2.577***	–2.756***	–0.891*
	(0.813)	(0.811)	(0.825)	(0.531)
Retail	–5.157***	–4.909***	–4.968***	–3.087***
	(0.409)	(0.408)	(0.412)	(0.285)
Finance	4.393***	4.272***	4.258***	3.223***
	(0.594)	(0.592)	(0.605)	(0.372)
Education	–0.881***	–0.811***	–0.891***	–0.170
	(0.207)	(0.206)	(0.211)	(0.126)
Administration	0.787**	0.682**	0.713**	0.011
	(0.309)	(0.308)	(0.318)	(0.177)
Armed forces	–2.170***	–1.985***	–2.038***	–0.892**
	(0.632)	(0.630)	(0.635)	(0.448)
ρ	–	0.084***	–	0.754***
(Spatial autocorrelation)		(0.011)		(0.007)
λ	–	–	0.082***	–0.946***
(Spatial error dependency)			(0.011)	(0.013)
Time FE	Yes	Yes	Yes	Yes
Loglikelihood	–37,893.480	–37,847.430	–37,851.660	–37,133.580
AIC	75,816.960	75,728.850	75,737.330	74,303.150
Observation	21,931	21,931	21,931	21,931

B: Measures of spatial effects

	Direct effect	Indirect effect	Total effect
ln(Population)	0.099	0.240	0.338
ln(Income)	0.067	0.162	0.229
Wholesale	–1.056	–2.568	–3.624
Retail	–3.657	–8.895	–12.552
Finance	3.818	9.287	13.105
Education	–0.202	–0.491	–0.693
Administration	0.013	0.033	0.046
Armed forces	–1.057	–2.571	–3.628

Note: We take the transformation of natural logarithm for two continuous variables (population and income). "FE" stands for the fixed effect; *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

One heterogeneous outcome from the case of general cyber risks in Section 5.2 is that a spatial coefficient (ρ) is positive, implying that a county with large population or high income and its neighboring regions are more likely to be exposed to data breaches. However, Table C3 still supports the main results of loss frequency that the financial industry variable has a positive coefficient, but the other industry variables are negative. A positive spatial coefficient leads to the results that counties with higher proportion of the financial industry and public administration sector and those neighboring regions tend to be more exposed to data breaches. In contrast, counties with more businesses in the other four industries and those adjacent regions tend to be less exposed to data breach events.

Table C4 presents the results of four models for three aspects examining socioeconomic features of data breach events in hot clusters (California and the Northeast Coast area), and Table C5 shows the results of two risk types (hacking and non-

TABLE C4
Results of Spatial Panel Analysis in Cases of the Two Most Exposed Regions

A: Comparison of regression results

	Data breach loss frequency in California				Data breach loss frequency in the Northeast Coast area			
	SLM	SAR	SEM	SAC	SLM	SAR	SEM	SAC
Constant	–35.823***	–36.454***	–35.533***	–36.272***	4.973	7.371	6.387	–17.627
	(8.865)	(8.698)	(8.698)	(8.698)	(16.552)	(16.367)	(16.378)	(15.219)
ln(Population)	1.935***	1.944***	1.932***	1.942***	–0.148	–0.097	–0.138	0.156
	(0.132)	(0.133)	(0.129)	(0.140)	(0.287)	(0.287)	(0.284)	(0.288)
ln(Income)	2.717***	2.771***	2.680***	2.752***	0.264	0.002	0.120	2.296
	(0.833)	(0.840)	(0.813)	(0.877)	(1.548)	(1.531)	(1.526)	(1.584)
Wholesale	–118.677***	–117.541***	–118.692***	–117.763***	–149.748***	–150.720***	–149.736***	–133.643***
	(22.149)	(21.951)	(21.708)	(22.321)	(36.994)	(36.596)	(36.534)	(35.334)
Retail	–36.144***	–36.112***	–35.895***	–36.060***	–31.724*	–33.095*	–32.172*	–29.458*
	(8.666)	(8.505)	(8.494)	(8.508)	(17.428)	(17.240)	(17.214)	(16.826)
Finance	–43.195***	–43.036***	–42.069***	–42.805***	80.456***	81.147***	80.736***	66.698***
	(10.800)	(10.638)	(10.577)	(10.714)	(17.671)	(17.475)	(17.459)	(16.759)
Education	–19.477***	–19.637***	–19.298***	–19.567***	–7.545	–7.979	–7.571	–12.041*
	(4.937)	(4.891)	(4.840)	(4.975)	(6.797)	(6.734)	(6.720)	(6.352)
Administration	1.535	1.640	1.581	1.632	–15.693	–16.158	–15.770	–17.064
	(3.933)	(3.865)	(3.859)	(3.876)	(11.694)	(11.569)	(11.511)	(12.742)
Armed forces	–44.911**	–45.048**	–45.767**	–45.224**	18.001	17.880	18.070	10.202
	(18.253)	(17.919)	(17.902)	(17.936)	(21.588)	(21.351)	(21.348)	(20.129)
ρ	–	–0.021	–	–0.017	–	–0.041	–	–0.554***
		(0.060)		(0.100)		(0.052)		(0.100)
λ	–	–	–0.025	–0.006	–	–	–0.014	0.485***
			(0.075)	(0.125)			(0.054)	(0.076)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
LL	–992.148	–992.088	–992.098	–992.087	–2,279.115	–2,278.728	–2,279.077	–2,272.338
AIC	2,014.296	2,018.175	2,018.195	2,020.174	4,588.231	4,591.457	4,592.155	4,580.675
Observation	406	406	406	406	721	721	721	721

B: Measures of spatial effects

	California			New York		
	Direct effect	Indirect effect	Total effect	Direct effect	Indirect effect	Total effect
ln(Population)	–	–	–	0.166	–0.066	0.100
ln(Income)	–	–	–	2.448	–0.971	1.477
Wholesale	–	–	–	–142.518	56.508	–86.010
Retail	–	–	–	–31.414	12.456	–18.958
Finance	–	–	–	71.127	–28.202	42.926
Education	–	–	–	–12.840	5.091	–7.749
Administration	–	–	–	–18.197	7.215	–10.982
Armed forces	–	–	–	10.880	–4.314	6.566

Note: We take the transformation of natural logarithm for two continuous variables (population and income). "FE" stands for the fixed effect; *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively. Panel B of the California case does not provide the results because no evidence on spatial effects is observed.

TABLE C5
Results of Spatial Panel Analysis in Cases of Two Risk Classes

A: Comparison of regression results

	Data breach loss frequency by HACK				Data breach loss frequency by NON-HACK			
	SLM	SAR	SEM	SAC	SLMe	SAR	SEM	SAC
Constant	–3.059***	–2.865***	–3.105***	–0.634***	–2.528***	–2.366***	–2.578***	–0.536**
	(0.279)	(0.278)	(0.278)	(0.231)	(0.259)	(0.258)	(0.258)	(0.223)
ln(Population)	0.097***	0.091***	0.095***	0.035***	0.125***	0.118***	0.124***	0.050***
	(0.004)	(0.04)	(0.004)	(0.002)	(0.004)	(0.004)	(0.004)	(0.002)
ln(Income)	0.242***	0.227***	0.249***	0.042***	0.162***	0.151***	0.168***	0.017
	(0.027)	(0.027)	(0.029)	(0.015)	(0.026)	(0.026)	(0.026)	(0.015)
Wholesale	–1.630***	–1.635***	–1.720***	–0.703**	–0.872**	–0.944**	–1.046**	–0.193
	(0.450)	(0.449)	(0.457)	(0.290)	(0.418)	(0.417)	(0.424)	(0.285)
Retail	–2.408***	–2.279***	–2.301***	–1.419***	–2.749***	–2.622***	–2.659***	–1.714***
	(0.227)	(0.226)	(0.228)	(0.155)	(0.210)	(0.210)	(0.212)	(0.153)
Finance	1.994***	1.946***	1.945***	1.414***	2.398***	2.322***	2.309***	1.844***
	(0.329)	(0.328)	(0.336)	(0.203)	(0.306)	(0.305)	(0.312)	(0.200)
Education	–0.463***	–0.434***	–0.475***	–0.110	–0.417***	–0.375***	–0.416***	–0.069
	(0.115)	(0.114)	(0.117)	(0.069)	(0.106)	(0.106)	(0.109)	(0.068)
Administration	0.373**	0.319*	0.337*	–0.022	0.415***	0.360**	0.372**	0.049
	(0.171)	(0.170)	(0.176)	(0.097)	(0.159)	(0.158)	(0.164)	(0.096)
Armed forces	–1.165***	–1.070***	–1.094***	–0.518**	–1.004***	–0.909***	–0.939***	–0.397*
	(0.350)	(0.349)	(0.352)	(0.245)	(0.325)	(0.324)	(0.327)	(0.239)
ρ	–	0.088***	–	0.767***	–	0.086***	–	0.722***
		(0.011)		(0.007)		(0.011)		(0.008)
λ	–	–	0.087***	–0.962**	–	–	0.085***	–0.887***
			(0.011)	(0.013)			(0.011)	(0.015)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
LL	–24,935.670	–24,885.310	–24,888.940	–24,093.380	–23,312.130	–23,266.480	–23,269.970	–22,746.860
AIC	49,901.340	49,804.630	49,811.890	48,222.760	46,654.270	46,566.950	46,573.940	45,529.710
Observation	21,931	21,931	21,931	21,931	21,931	21,931	21,931	21,931

B: Measures of spatial effects

	HACK			NON-HACK		
	Direct effect	Indirect effect	Total effect	Direct effect	Indirect effect	Total effect
ln(Population)	0.042	0.110	0.152	0.059	0.123	0.182
ln(Income)	0.051	0.131	0.182	0.019	0.041	0.060
Wholesale	–0.841	–2.179	–3.020	–0.224	–0.471	–0.695
Retail	–1.697	–4.398	–6.095	–1.988	–4.183	–6.172
Finance	1.691	4.382	6.074	2.138	4.499	6.637
Education	–0.132	–0.341	–0.473	–0.080	–0.169	–0.250
Administration	–0.027	–0.070	–0.097	0.056	0.119	0.175
Armed forces	–0.619	–1.604	–2.223	–0.460	–0.968	–1.428

Note: We take the transformation of natural logarithm for two continuous variables (population and income). "FE" stands for the fixed effect; *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

hacking types). The risk classification of the PRC is simpler than that of the Advisen dataset (see panel B of Table C1), so we simplify this aspect into two risk types.[43] The hacking type can be aligned with the malicious and privacy types of Table 7, whereas the nonhacking type can represent the unintended risk type in the main result. In both Tables C4 and C5, we overall find consistent consequences with the main results in terms of two key socioeconomic factors (population and average income) that are both positive and significant in addressing loss frequency, and with the significance of the income variable in the optimal model (SAC).

It is also observed in the PRC case that the impacts of the financial industry variable are heterogeneous between two regions, where it is negative (positive) and significant for the California (the Northeast Coast) region. Thus, the narrow scope with only data breach risks (as part of cyber risks) also supports industrial heterogeneity in risk exposures of hot clusters. Moreover, we find no evidence of key variables (population and income level) in the Northeast Coast case and the spatial effect in the California case. However, except for the California case the results of this data breach case overall support our main results with general cyber risks in terms of the other industrial variables and spatial effects (direct and indirect effects).

APPENDIX D. SPATIAL REGRESSIONS ON OTHER STATES

As a consistency test, we further investigate two states (Texas and Florida) representing regions with high weight in the medical industry and education.[44] In Tables D1 and D2, we find mostly consistent results with those of California and the Northeast Coast region. Population as the key factor representing the size of exposures has overall positive, significant effects on both loss frequency and severity of the two states. This result implies that regions with relatively dense population and presumably a large number of Internet users might be more exposed to cyber risks. Mixed results are observed for

TABLE D1
Results of Spatial Panel Analysis in the State of Texas

A: Comparison of regression results

	Cyber loss frequency				Cyber loss severity			
	SLM	SAR	SEM	SAC	SLM	SAR	SEM	SAC
Constant	–29.311***	–29.138***	–30.428***	–26.014**	–26.224	–25.950***	–27.269***	–32.058***
	(10.389)	(10.352)	(10.348)	(10.314)	(2.695)	(2.685)	(2.674)	(2.631)
ln(Population)	2.627***	2.624***	2.624***	2.467***	1.038***	1.031***	1.053***	1.136***
	(0.128)	(0.132)	(0.125)	(0.157)	(0.033)	(0.035)	(0.034)	(0.038)
ln(Income)	1.352	1.337	1.453	1.109	1.770***	1.748***	1.863***	2.291***
	(0.999)	(1.003)	(0.987)	(1.004)	(0.259)	(0.262)	(0.264)	(0.278)
Wholesale	–18.287	–18.305	–18.372	–19.173	–2.428	–2.434	–2.153	–1.420
	(11.897)	(11.857)	(11.819)	(11.676)	(3.086)	(3.075)	(3.086)	(3.057)
Retail	–50.906***	–50.863***	–51.806***	–52.290***	–12.308***	–12.245***	–11.578***	–11.064***
	(6.735)	(6.712)	(6.698)	(6.645)	(1.747)	(1.741)	(1.740)	(1.711)
Finance	66.296***	66.198***	69.163***	72.632***	18.916***	18.767***	17.336***	16.208***
	(10.864)	(10.825)	(10.786)	(10.647)	(2.818)	(2.808)	(2.822)	(2.798)
Education	–25.801***	–25.792***	–25.424***	–23.867***	–2.433**	–2.400**	–2.551**	–3.167***
	(3.948)	(3.943)	(3.914)	(3.922)	(1.024)	(1.023)	(1.031)	(1.039)

(*Continued*)

[43]As for the Advisen case of the main study, we again exclude the unknown risk type, which may lead to a sample bias due to lack of information on risk characteristics (Eling and Jung 2018).

[44]We refer to the 2023 statistics from American Hospital Directory (AHD) and the 2022 statistics provided by National Center for Education Statistics (NCES), where California, Texas, and Florida are the states with the largest numbers of hospitals and higher education institutions (AHD 2023; NCES 2024).

TABLE D1
(*Continued*)

A: Comparison of regression results

	Cyber loss frequency				Cyber loss severity			
	SLM	SAR	SEM	SAC	SLM	SAR	SEM	SAC
Administration	2.630	2.602	2.953	2.643	3.788**	3.727**	3.440**	3.487**
	(5.871)	(5.852)	(5.814)	(5.710)	(1.523)	(1.519)	(1.538)	(1.552)
Armed forces	–55.949***	–55.828***	–56.534***	–52.584***	2.533	2.634	1.379	–1.575
	(16.160)	(16.145)	(16.016)	(16.033)	(4.192)	(4.189)	(4.226)	(4.263)
ρ	–	0.002	–	0.105	–	0.016	–	–0.221***
		(0.030)		(0.071)		(0.027)		(0.051)
λ	–	–	–0.036	–0.143*	–	–	0.121***	0.311***
			(0.033)	(0.084)			(0.031)	(0.049)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
LL	–9,161.859	–9,161.856	9,161.325	–9,159.627	–5,734.341	–5,734.168	–5,727.454	–5,717.374
AIC	18,359.718	18,363.712	18,362.650	18,361.254	11,504.682	11,508.336	11,494.908	11,476.748
Observation	2,540	2,540	2,540	2,540	2,540	2,540	2,540	2,540

B: Measures of spatial effects

	Direct effect	Indirect effect	Total effect	Direct effect	Indirect effect	Total effect
ln(Population)	–	–	–	1.151	–0.220	0.931
ln(Income)	–	–	–	2.321	–0.444	1.877
Wholesale	–	–	–	–1.439	0.275	–1.163
Retail	–	–	–	–11.209	2.144	–9.065
Finance	–	–	–	16.420	–3.141	13.279
Education	–	–	–	–3.209	0.614	–2.595
Administration	–	–	–	3.533	–0.676	2.857
Armed forces	–	–	–	–1.596	0.305	–1.291

Note: We take the transformation of natural logarithm for two continuous variables (population and income). "FE" stands for the fixed effect; *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

TABLE D2
Results of Spatial Panel Analysis in the State of Florida

A: Comparison of regression results

	Cyber loss frequency				Cyber loss severity			
	SLM	SAR	SEM	SAC	SLM	SAR	SEM	SAC
Constant	99.362***	113.286***	113.346***	109.206***	–31.663***	–33.240***	–33.376***	–33.610***
	(28.374)	(27.755)	(27.074)	(26.975)	(7.807)	(7.679)	(7.673)	(7.675)
ln(Population)	9.335***	9.185***	9.741***	9.784***	3.137***	3.208***	3.158***	3.191***
	(0.530)	(0.526)	(0.518)	(0.519)	(0.146)	(0.147)	(0.142)	(0.153)
ln(Income)	–18.289***	–19.970***	–20.672***	–20.195***	0.566	0.750	0.790	0.804
	(2.944)	(2.891)	(2.751)	(2.832)	(0.810)	(0.799)	(0.800)	(0.803)
Wholesale	96.186*	73.985	70.589	75.077	–27.143*	–25.006	–28.218*	–26.697*
	(57.832)	(56.817)	(55.888)	(56.166)	(15.912)	(15.711)	(15.489)	(15.695)

(*Continued*)

TABLE D2
(*Continued*)

A: Comparison of regression results

	Cyber loss frequency				Cyber loss severity			
	SLM	SAR	SEM	SAC	SLM	SAR	SEM	SAC
Retail	–156.592***	–142.276***	–135.774***	–137.971***	–37.975***	–40.048***	–40.191***	–40.498***
	(22.499)	(22.214)	(21.481)	(21.701)	(6.190)	(6.123)	(6.069)	(6.169)
Finance	150.659***	158.390***	171.447***	169.865***	–8.315	–10.053	–11.165	–11.080
	(31.200)	(30.520)	(30.954)	(31.085)	(8.584)	(8.452)	(8.285)	(8.367)
Education	–55.197***	–47.293***	–46.801***	–48.601***	–8.031**	–9.935***	–9.650***	–10.073***
	(12.099)	(11.970)	(12.235)	(12.606)	(3.329)	(3.353)	(3.201)	(3.415)
Administration	60.645***	65.196***	61.705***	59.584***	15.772***	14.865***	16.624***	15.869***
	(14.415)	(14.165)	(15.018)	(15.475)	(3.966)	(3.969)	(3.782)	(3.971)
Armed forces	–198.787***	–181.159***	–170.932***	–171.420***	–44.176***	–48.215***	–50.656***	–50.529***
	(49.156)	(48.274)	(52.345)	(52.724)	(13.525)	(13.315)	(12.721)	(12.985)
ρ	–	0.147***	–	–0.058	–	–0.081*	–	–0.043
		(0.044)		(0.082)		(0.044)		(0.068)
λ	–	–	0.303***	0.336***	–	–	–0.113*	–0.071
			(0.049)	(0.081)			(0.058)	(0.089)
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
LL	–2,560.162	–2,556.154	–2,545.282	–2,544.918	–1,695.548	–1,694.071	–1,693.971	–1,693.778
AIC	5,156.325	5,152.308	5,130.564	5,131.836	3,427.097	3,428.143	3,427.943	3,429.539
Observation	670	670	670	670	670	670	670	670

B: Measures of spatial effects

	Direct effect	Indirect effect	Total effect	Direct effect	Indirect effect	Total effect
ln(Population)	–	–	–	–	–	–
ln(Income)	–	–	–	–	–	–
Wholesale	–	–	–	–	–	–
Retail	–	–	–	–	–	–
Finance	–	–	–	–	–	–
Education	–	–	–	–	–	–
Administration	–	–	–	–	–	–
Armed forces	–	–	–	–	–	–

Note: We take the transformation of natural logarithm for two continuous variables (population and income). "FE" stands for the fixed effect; *, **, and *** indicate significance levels of 10%, 5%, and 1%, respectively.

the average income level, being positive and partially significant in the case of Texas, but either negative or with insignificant effects in the frequency model of Florida.

These mixed observations may result from the finding that California and the Northeast Coast are found to be red spots with high-risk exposures in Figure 3, whereas the states of Texas and Florida are not, thereby possibly leading to some differences in the statistical results in between. Furthermore, unlike California and the Northeast Coast, most cases except for the severity model of Texas show no spatial effects. This result may imply that compared to California and the Northeast Coast, which are known for large-scale financial hubs and technology centers (e.g., Silicon Valley), states with large-sized medical and educational industries might be likely to face risks that are less clustered (i.e., less concentrated).

APPENDIX E. LOG-LIKELIHOOD RATIO TEST AND GRAPHICAL DIAGNOSIS

TABLE E1
Log-Likelihood Ratio Test between the SLM and the Optimal Model by Regions

	Loss frequency		Loss severity	
	SLM	Optimal model	SLM	Optimal model
A: Log-likelihood ratio test for all counties				
Log-likelihood	–122,044.500	–120,931.300	–74,166.270	–73,998.220
LLR test	2,226.400***		336.100***	
B: Log-likelihood ratio test for California				
Log-likelihood	–2,877.422	–2,859.337	–1,519.155	–1,507.178
LLR test	36.170***		23.954***	
C: Log-likelihood ratio test for Northeast Coast				
Log-likelihood	–4,993.652	–4,944.904	–2,610.838	–2,606.795
LLR test	97.496***		8.086***	

Note: The LLR stands for log-likelihood ratio. For the LLR test, we present the test statistic and significance level with *, **, and *** indicating 10%, 5%, and 1%, respectively.

TABLE E2
Log-Likelihood Ratio Test between the SLM and the Optimal Model by Risk Type

	Loss frequency		Loss severity	
	SLM	Optimal model	SLM	Optimal model
A: Log-likelihood ratio test for malicious risk type				
Log-likelihood	–100,217.20	–98,525.23	–72,147.06	–71,984.35
LLR test	3,383.940***		325.420***	
B: Log-likelihood ratio test for unintended risk type				
Log-likelihood	–74,696.70	–73,324.33	–64,511.13	–64,376.79
LLR test	2,744.740***		268.680***	
C: Log-likelihood ratio test for privacy risk type				
Log-likelihood	–99,643.37	–99,059.27	–59,320.47	–58,724.77
LLR test	1,168.200***		1,191.400***	

Note: The LLR stands for log-likelihood ratio. For the LLR test, we present the test statistic and significance level with *, **, and *** indicating 10%, 5%, and 1%, respectively.

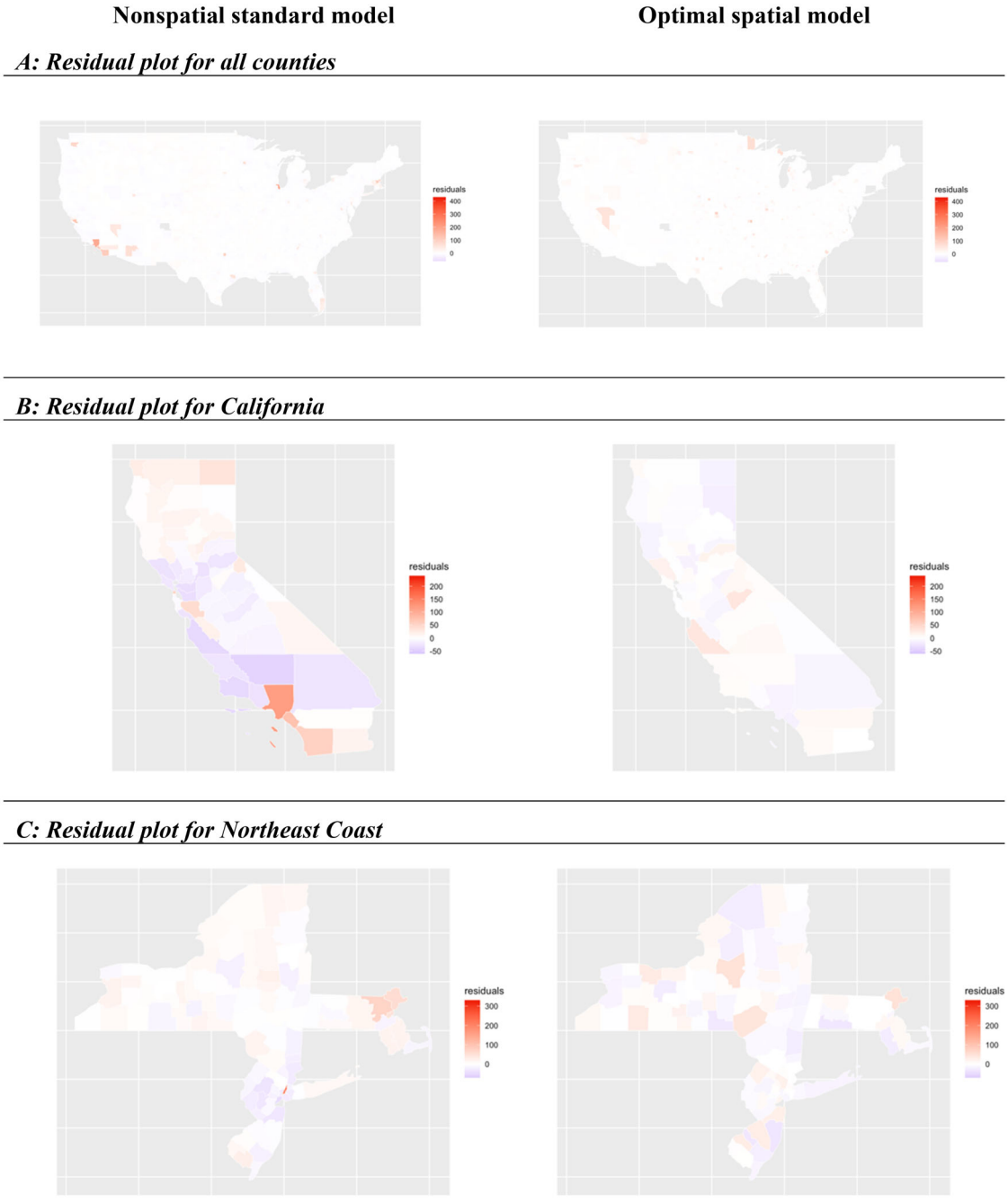

FIGURE E1. Residual Plots of Nonspatial Standard Model and the Optimal Spatial Model for Loss Frequency (Three Regional Categories).

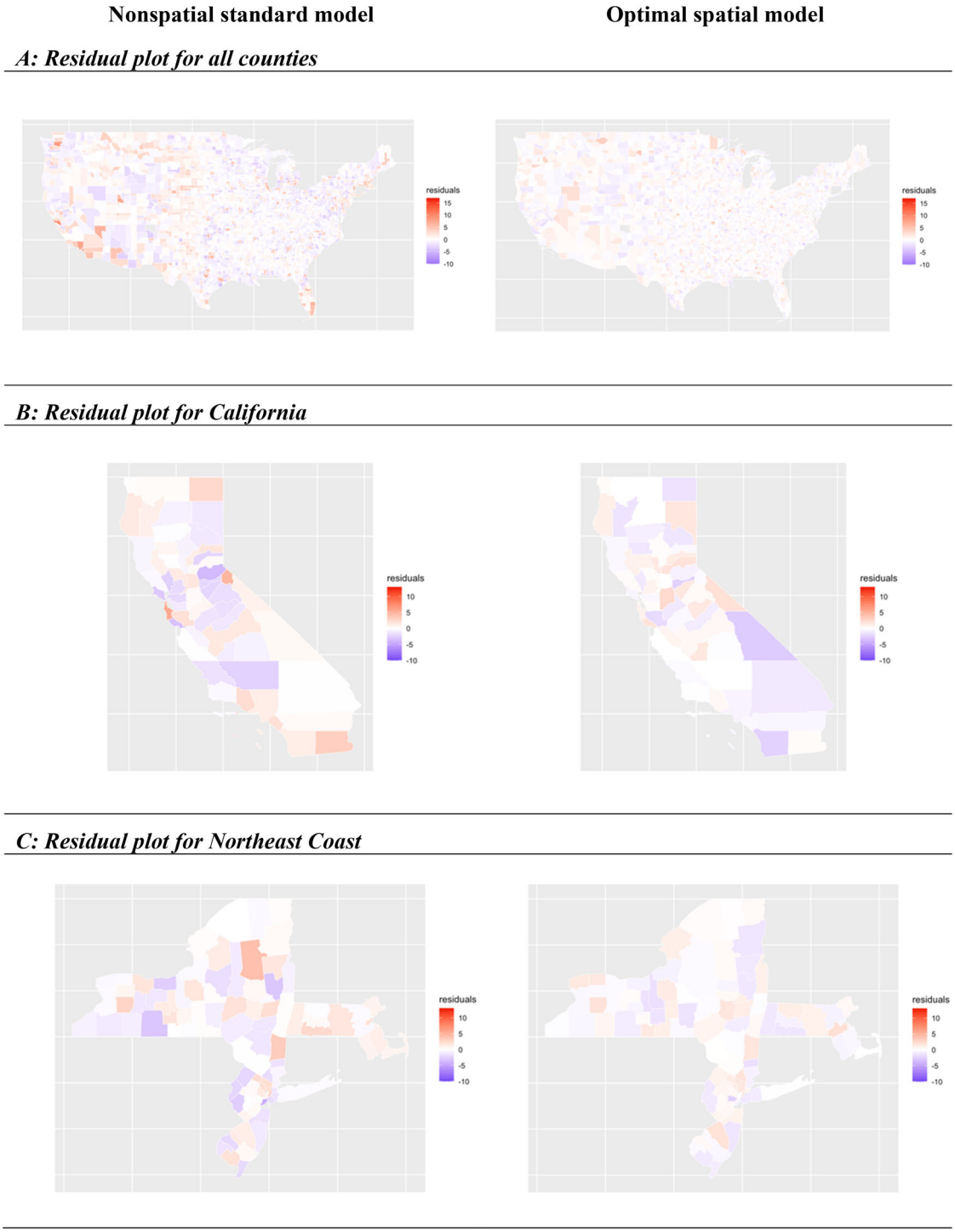

FIGURE E2. Residual Plots of Nonspatial Standard Model and the Optimal Spatial Model for Loss Severity (Three Regional Categories).

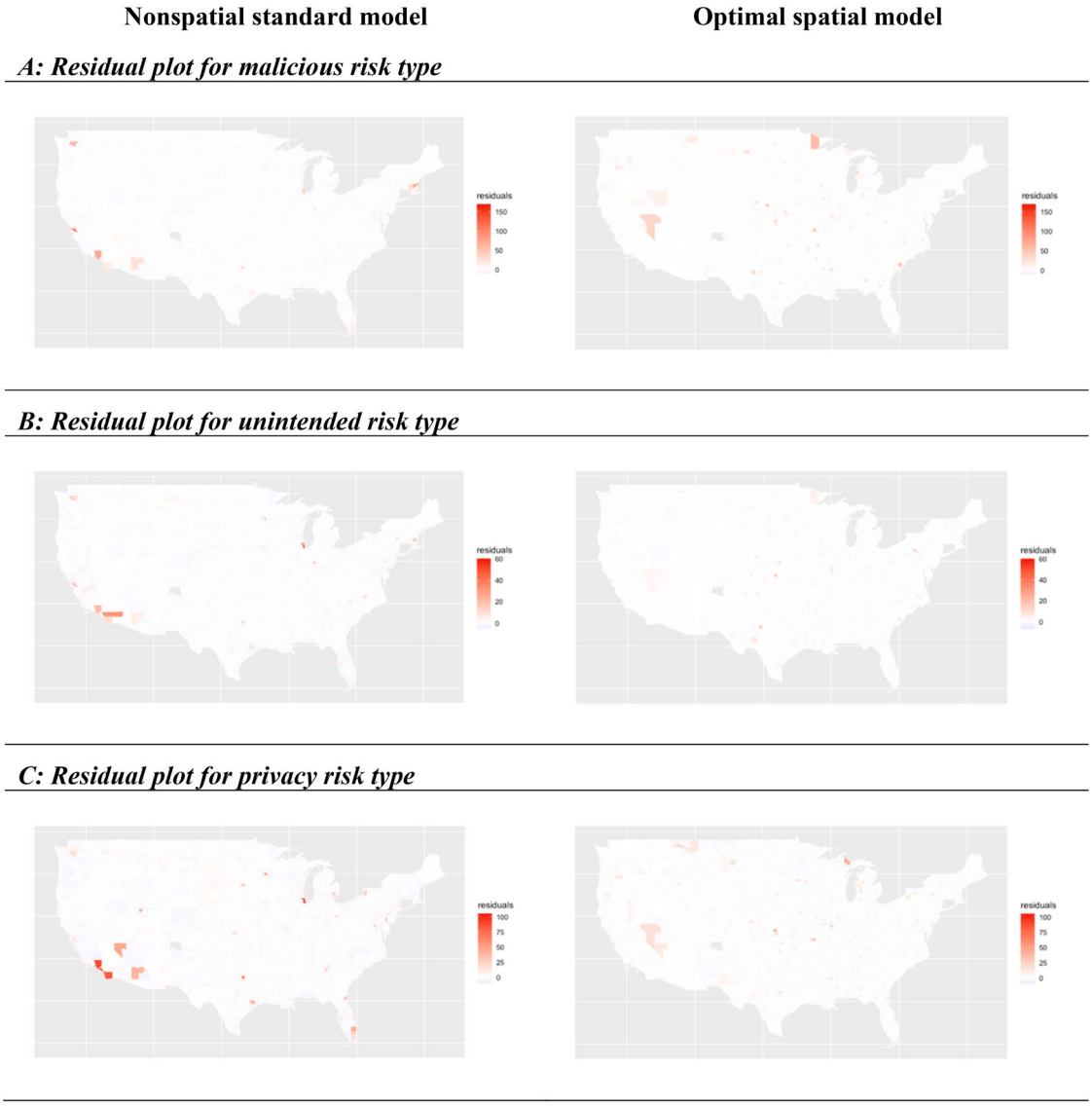

FIGURE E3. Residual Plots of Nonspatial Standard Model and the Optimal Spatial Model for Loss Frequency (Three Risk Types).

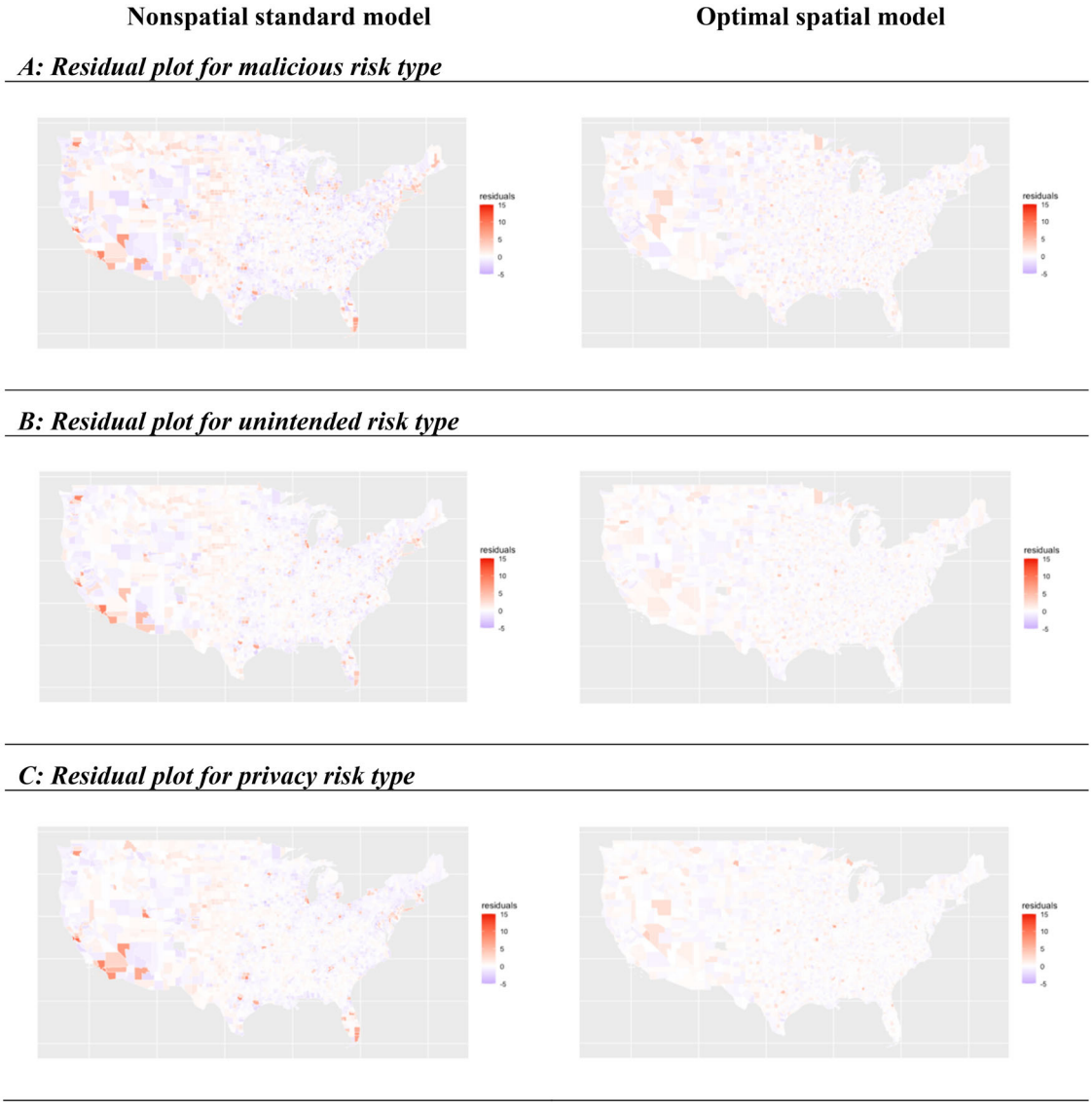

FIGURE E4. Residual Plots of Nonspatial Standard Model and the Optimal Spatial Model for Loss Severity (Three Risk Types).

REFERENCES FOR THE APPENDIX

American Hospital Directory (AHD). 2023. *Hospital statistics by state.* https://www.ahd.com/state_statistics.html.

Florackis, C., C. Louca, R. Michaely, and M. Weber. 2023. Cybersecurity risk. *Review of Financial Studies* 36 (1):351–407. doi:10.1093/rfs/hhac024

Kamiya, S., J. K. Kang, J. Kim, A. Milidonis, and R. M. Stulz. 2021. Risk management, firm reputation, and the impact of successful cyberattacks on target firms. *Journal of Financial Economics* 139 (3):719–49. doi:10.1016/j.jfineco.2019.05.019

National Center for Education Statistics (NCES). 2024. *Digest of education statistics: 2022.* https://nces.ed.gov/programs/digest/d23/tables/dt23_317.20.asp.

National Conference of State Legislatures (NCSL). 2022. Security breach notification laws. https://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx.