Summer 6-15-2016

AN OPEN DOOR MAY TEMPT A SAINT – DATA ANALYTICS FOR SPATIAL CRIMINOLOGY

Johannes Bendler
University of Freiburg, johannes.bendler@is.uni-freiburg.de

Tobias Brandt
University of Freiburg, tobias.brandt@is.uni-freiburg.de

Dirk Neumann
University of Freiburg, dirk.neumann@is.uni-freiburg.de

AN OPEN DOOR MAY TEMPT A SAINT – DATA ANALYTICS FOR SPATIAL CRIMINOLOGY

Research

Bendler, Johannes, University of Freiburg, Freiburg, Germany, johannes.bendler@is.uni-freiburg.de

Brandt, Tobias, University of Freiburg, Freiburg, Germany, tobias.brandt@is.uni-freiburg.de

Neumann, Dirk, University of Freiburg, Freiburg, Germany, dirk.neumann@is.uni-freiburg.de

Abstract

The vast amounts of data that are generated and collected in today's world bear immense potential for businesses and authorities. Innovative companies already adopt novel analytics methods driven by competition and the urge of constantly gaining new insights into business operations, customer preferences, and strategic decision making. Nonetheless, local authorities have been slow to embrace the opportunities enabled by data analytics. In this paper, we demonstrate and discuss how latent structures unveil valuable information on an aspect of public life and communities we all face: criminal activity. On city-scale, we analyze the spatial correspondence of recorded crime to its physical environment, the public presence, and the demographical structure in its vicinity. Our results show that Big Data in fact is able to identify and quantify the main spatial drivers of criminal activity. At the same time, we are able to maintain interpretability by design, which ultimately allows deep informational insights.

Keywords: Data Analytics, Social Media, Spatial Data, PLS, Criminal Activity

1 Introduction

In recent years, businesses have begun to seize the opportunities created by the vast amounts of data generated and collected in today's world. Business Intelligence and Big Data Analytics provide novel insights into business operations, customer preferences, and strategic decision-making. At the same time, the use of information technology systems in the public sector continues to increase. This development is exemplified by, for instance, e-government initiatives that seek to simplify interactions between citizens and public administrations and to facilitate improved participation in democratic decision-making processes (Bertot et al. 2010; Effing et al. 2011; Irvin and Stansbury 2004).

However, compared to the business world, the public sector has been slow to embrace the opportunities enabled by the Big Data paradigm. Messatfa et al. (2011) summarize that "most public organization are just starting to explore ways to leverage analytics" (p. 2). The reasons for this reluctance are manifold and may include strict hierarchies, a lack of competition, and data privacy restrictions. Incidentally, public administrations and organizations fundamentally rely on a certain trust from the general population and may limit themselves in handling and analyzing data to a degree even exceeding legal requirements. To provide the public with perspectives on the tradeoff between data privacy and improved public services, researchers need to identify potential benefits and risks associated with data analytics applications in the public sector.

In this paper, we demonstrate and discuss how novel data analytics methods may improve capabilities of agencies working on a crucial aspect of public life: combatting crime. The protection of its citizens and the enforcement of the law are two primary duties of the modern state. Hence, police departments seek to suppress crime while city planners and administrations want to identify drivers and causes of crime hot spots in particular neighborhoods. While such spatial evaluations of criminal incidence have been conducted for several decades (Bowes 2007; Brantingham and Brantingham 1993; Cohen 1941; Grubesic and Mack 2008; Ratcliffe 2004; Roncek 1981), the availability of new data sources and techniques for analysis and visualization provides novel insights. For instance, predictive policing uses historical data on criminal incidents to identify hot spots and the likelihood of future crimes in certain areas (Bachner 2013). The objective of this paper is to identify the underlying reasons of this spatial variation in criminal incidents—to determine why certain neighborhoods are more prone to crime than others—through spatial data analytics.

Over the past decade, the availability of data with a spatial component has substantially evolved (Bendler et al. 2014b). In these days, even smartphones contain a GPS-unit for routing that also enables social media platforms and recommender systems to capture and share the position of the user. Map services, such as Google Maps or OpenStreetMap, provide comprehensive information on the structural environment, along with details on businesses and attractions in particular areas. These data sources produce novel explanatory variables for spatial phenomena. However, they are also affected by a central critique of Big Data approaches—a lack of theoretical foundation. Chang et al. (2014) argue that analytics methods in the Big Data era continue to require a basis in theory to result in meaningful insights. For instance, in the context of criminal incidents, an analysis may show that a higher number of cafés may significantly correlate with the occurrence of thefts in the vicinity. However, this direct correlation does not necessarily imply a direct relationship since there is no theoretical argument for café owners or patrons being particularly prone to crime. Neither does opening a new café automatically results in a surge of thefts. Instead, the number of crimes may, together with other variables, reflect an unobservable, latent characteristic of the area. These characteristics, if discovered, can be compared to theories on the determinants of crime.

We employ the Projection to Latent Structures (PLS) method to identify latent characteristics that explain the spatial variation in criminal incidents and that can be supported by a basis in theory. Analyzing a data set comprised of several sources—Twitter, the US Census, police departments, and map services—we combine novel spatial analytics methods with a methodology that is well established in Information Systems research. We, thereby, outline how data analytics can improve a public service that is crucial to the functioning of modern societies. For this purpose, we address the following research questions during the course of this paper.

(1) How can Big Data Analytics support the identification of crucial characteristics that explain the spatial variation of crime?

(2) Can the identified latent structures be related to theoretical concepts, thus improving the interpretability of the results?

In the following section, we provide an overview of the relevant literature concerning spatial criminology and projection to latent structures. Subsequently, we present our research outline for this paper describing the sequence we use to employ PLS methodology, regressions, and prediction. The two succeeding sections provide detailed explanations of the applied methodology. The gained insights are then fed into a predictive approach to forecast spatial criminal activity. Finally, we discuss our results and provide a prospect on future research opportunities.

2 Related work

The investigation of spatial and temporal characteristics of crime has been an active field of research for some time. Over the years, many different approaches have emerged that aim to explain or predict patterns of criminal activity in space and time. Such crime-related studies cover a broad range of demographic, structural, environmental, and behavioral factors and research their respective impact on the emergence of various different crime types.

2.1 Spatial Criminology

Especially the spatial investigation received increased attention throughout the time. Cohen (1941) states that, in the forty years prior to his publication, "students have become less and less disposed […] concerning the influences of physical geography upon crime" (p. 29). Starting from the mid-20th century, research on criminology was however carried out with respect to a broad variety of possible influencing aspects, especially focusing on socio-spatial characteristics. Block (1979) investigates on homicide, robbery, and aggravated assault crimes and states that, "using regression analysis […] [it] is found that neighborhoods in which very poor and middle-class people live in close proximity are those in which rates of all three types of criminal violence are highest" (p. 46). Roncek (1981) traces criminal activity back to three major hypotheses (1) household composition, (2) features of the residential environment, and (3) the interaction of the social composition and the features of the residential environment. By introduction of data on household composition and social components for crime prediction, Roncek's hypotheses lay the foundation of what current research still includes in criminological studies—the relationship between criminal incidents, a delinquent's social background, and the corresponding geographic environment. Besides the pure demographic milieu, the (partially corresponding) levels of wealth and education are being considered (Patterson 1991). Nevertheless, researchers have mostly studied the direct effects of sociographic, demographic, and geographic measures on the emergence of crimes in general or selected types of crime. Contrastingly, Brantingham and Brantingham (1993) refrain from searching for a direct association and state that "crime has long been thought to be intimately associated with the physical environment in which it occurs. […] The relationship between crime and the physical environment is mediated through individual awareness and action spaces" (p. 3). Still, the direct effects of physical environment reside as an active research field (Andresen 2006; Chainey et al. 2008; Krivo and Peterson 1996; Murray 2001; Nelson et al. 2001).

Starting in the mid-90s, computer systems improved in performance and storage and thereby started to render more complex analyses possible. Novel tools from that era, such as Geographic Information Systems (GIS), were employed to deepen the insights of spatial criminology (e.g. Bowers 1999) Driven by the increased availability of data roughly starting with the beginning of the 21st century, more complex models and calculations were carried out in the past decade (e.g. Ratcliffe 2004, 2010, Bows 2007, Grubesic and Mack 2008, Day 2014).

Recently, the availability of user data from online social interactions on various platforms (such as Twitter) permits the next step in criminological investigation. So far, the demographic and sociographic background of offenders, as well as the structural geographic environment are data sources that can reflect the intrinsic driving forces to criminal activity in certain regions. By now, we can also include

data that represents the public presence and possibly public attention by also including geo-located data from social networks. Traunmueller et al. (2014) have identified that previous work suggests that "there is a strong relationship between the built environment and location of crime" (p. 398). Both the initial work of Wang et al. (2012), as well as the recent study from Gerber (2014) confirm this perception. They provide an approach of predicting criminal activity by employing Twitter messages as a novel data source to reflect the location and mood of people in the vicinity of crimes. Similarly, Bendler et al. (2014a) employ Twitter data to improve the explanation of crimes in a spatial context. They identify several crime types that can be better explained by additionally including social media data.

2.2 Projection to Latent Structures

The Projection to Latent Structures (PLS, also: Partial Least Squares) approach is a technique to calculate the optimal projection of input variables to a lower number of latent structures among them by setting up linear combinations. It has been widely applied in social sciences, especially in the domain of marketing and consumer research (Henseler et al. 2009), for example for the purpose of confirmatory factor analysis. In 1980, Bookstein (1980) has provided a comprehensive description of the PLS approach covering various algorithms. Even though PLS oftentimes allows to reduce the variable space to a lower dimension efficiently, it is being conversely discussed. The two major points of criticism refer to (1) missing statistical tests due to absence of assumptions on probability distributions, and (2) the loss of interpretability by linearly combining seemingly unrelated input variables. Anyhow, PLS is oftentimes used as a method to account for causality and endogeneity, even though it bears immense risks in statistical terms. Antonakis et al. (2014) provide 10 commandments of causal analysis and explicitly exclude PLS from the application for the purpose of causality, because its results cannot be tested for overidentifying restrictions.

These weaknesses however only apply when employing PLS for investigating on causality and endogeneity. In a recent research note, Zheng and Pavlou (2010) employ PLS to purely identify latent structures in order to feed them into a Bayes network in a subsequent step. Even though PLS is usually employed in confirmatory approaches, the authors point out that it can also be applied for exploratory purposes. In addition, the related technique PLS Path Modeling (PLS-PM) allows to study the dependencies among latent variables in formative or reflective models. The field of PLS and PLS-PM is under highly active research and evolution (Dijkstra 2010; Dijkstra and Henseler 2015; Henseler and Sarstedt 2013; Hulland et al. 2010; McIntosh et al. 2014; Vinzi et al. 2010; Wold et al. 2010).

2.3 Research Gap

In this research work, we extend the previous approaches in spatial criminology that focus on hotspot/cluster analysis (e.g. Ratcliffe 2004; Curman et al. 2015) and observable structures (e.g. Gerber 2014; Bendler et al. 2014a) by introducing the PLS methodology to explore latent structures among our input variables. Lee et al. (2011) propose to perform an initial PLS analysis prior to the application of further methods when there is no coherent theory available in advance. Consequently, we will perform a PLS analysis using spatial data on demographics, sociographics, physical environment, social media, and criminal activity in order to seek for meaningful latent structures that can be well interpreted and support in answering our research questions posed at the outset.

3 Research Outline

As Henseler et al. (2009) delineate, many researchers state that their studies' goals relate to the particular strengths of PLS path modeling. Furthermore, the authors describe that "[t]he most important motivations are exploration and prediction, as PLS path modeling is recommended in an early stage of theoretical development in order to test and validate exploratory models." (Henseler et al. 2009, p. 282). Following this rationale, we employ PLS and PLS path modeling for exploration and early interpretation in the special case of spatial observations.

Our research concept is set up as outlined in Figure 1. The spatial observations from our four data sources census, map, social media, and crime are spatially aligned and preprocessed in order to fit the needs of

an exploratory analysis. In the first analysis step, we employ a PLS approach to explore a number of latent structures and their interrelations. Subsequently, the identified structures are fed into a path model to further investigate dependencies among them. Additionally, we perform a step-wise regression that provides insights on the explanatory power of each latent structure when investigating the emergence of crime. In a final step, we combine the results from both methods to carry out a predictive approach on criminal activity.

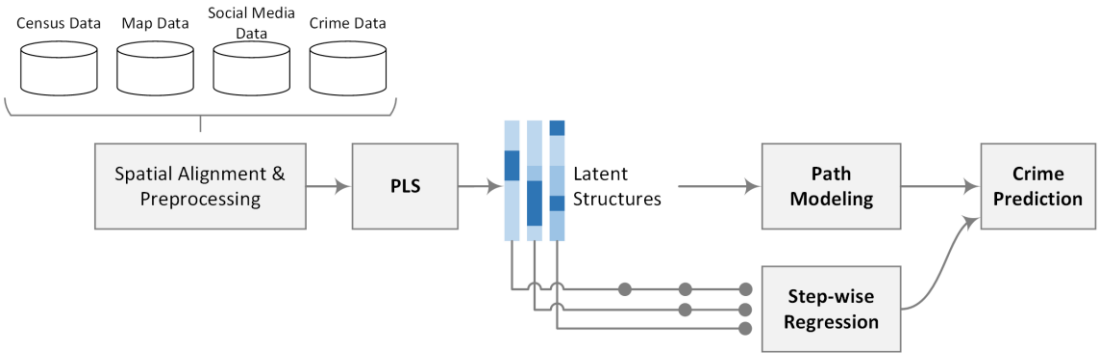


Figure 1. Research approach for latent structures in spatial criminology.

The aforementioned data sources provide measures and observations in various spatial references and resolutions. Thus, an efficient alignment and preprocessing is required to combine the data into a spatially consistent set of observations. In order to demonstrate our approach, we rely on the following data that refers to the geographical area of the City of San Francisco, whereby time-anchored data are collected during mid-2013, from June to August.

Census data reflects demographic and structural information on tract level and is freely obtainable via the US Census Bureau. Up to 2010, the US Census Bureau has collected complete census data (SF1) on a decennial base. Since then, the American Community Surveys (ACS), which provide the short-term changes in demographics, are published in shorter periods.

Map data refers to points of interest (POIs) as known from mapping services, such as OpenStreetMap, Google Maps, or Bing Maps. Since these POIs are provided in a large number of categories, we process them in clustered form. The resulting eight logical groups are authorities, entertainment, finance and law, food and drinks, health, retail and services, social and religion, and transportation.

Social media data is represented by over half a million geo-located Twitter messages from the months June to August in 2013. Each tweet contains exact coordinates in time and space, which allows to treat each of them as an indicator of a person in that spot at that very time. Therefore, we can expect the social media data to reflect public presence to a certain degree.

Crime data has become freely available from more and more urban areas over the past few years following the open data approach. We have obtained crime records as published by the SFPD covering 13 different crime types. Each delinquency record is accurately located in time and space. The 13 crime types are namely assault, burglary, disturbing the peace, drugs/alcohol violations, DUI, fraud, motor vehicle theft, robbery, sex crimes, theft/larceny, vandalism, vehicle break-in/theft, and weapons.

The structure of this research work is aligned to the steps described in Figure 1. The upcoming chapter deals with the preparation of our dataset and identifies latent structures among the input variables from our data. Subsequently, we assess the inner dependencies of the identified latent structures describing both path modeling and the step-wise regression approach. In the following chapter, we carry out a predictive analysis that employs insights from the previous chapters. This work closes with evaluation of our results and a chapter providing concluding remarks.

4 Projection to Latent Structures for Spatial Observations

In these times, the ubiquity of computing and, as a result, the commonly sensed “always-on” mentality generate masses of data, ranging from social interaction over online purchase to geographic traces of

individuals. The analytical potential is immense and opens the door for both businesses and governments to infer novel and detailed information on human behavior. Oftentimes, more data is said to be equivalent to more information. Contrastingly, extracting intrinsic information from Big Data is a matter of asking the right questions and heavily depends on interpretability of the data at hand. This principle also holds as valid when exploring and explaining spatio-temporal observations—seeking expressive interrelations and dependencies is subject to searching for a presumably meaningful combination of measures.

Especially in the domain of spatial analytics, researchers have to cope with the potentially high degree of multicollinearity and unknown interplay effects among different data sources. Due to the nature of urbanity, observations in fact are present in tight spatial relationship. For instance, a city center where stores and businesses are denser than in residential areas naturally also relates to a higher public presence during daytime, potentially higher housing prices, and an increased number of, for example, pickpocket incidents. These circumstances restrain analyzing the direct impact of spatial measures on to-be-explored observations. Latent structures can be employed as a hidden mid-layer that collects the effects of input measures (i.e. different data sources) and combines them according to their respective impact to joint nodes. Using historical data, we can calculate the likely impact of each data source on such latent structures and use the respective loadings to interpret the interplay and to quantify the impact.

As for our analysis, we mainly rely on the PLS methodology, which is superior over other approaches in terms of theoretical embedding. This corresponds to early interpretation and sense-making from unknown data in structural aspects. In general, PLS is able to project input variables to a much lower dimensional space by providing linear combinations among them and at the same time maintaining the maximum explanatory power. Zheng and Pavlou (2010) propose to use PLS to identify latent variables prior to feeding them into further steps of analysis—in their case a Bayesian approach.

4.1 Spatial Alignment and Preprocessing

In the setting of this research work, all observations are spatially anchored by design. As outlined before, spatial observations emerge in geographical clusters by the nature of cities. Tobler (1970) formulates his first law of geography by stating that everything is related to everything else, but near things are more related than distant things. Essentially, this means that a relationship between different spatial observations is possibly present despite a distance between them, and that this distance is the elemental factor that quantifies the relationship's strength. As a consequence, it is indispensable to prepare the spatial data to account for such distance-based relationships before entering the analysis itself.

While census data is available on tract level—represented by spatial polygons—all other sources provide observations as point data. The naïve approach of counting points in polygons does not account for the spatial dependencies outlined by Tobler and thus is inapplicable. In lieu thereof, we perform kernel density estimations (KDE) to transform point data to an area representation. The result of a KDE is comparable to a common heatmap and therefore simultaneously contains a distance weighting scheme.

Figure 2 delineates the alignment process that is applied to all spatial point observations using exemplary data. In a first step, the point data—shown in panel (a)—is blurred by application of a KDE (b). Using a sampling grid (c) of the desired resolution, we then can transform the continuous KDE result back to discrete points (d). For census data, which is available on a polygonal basis, only steps (c) and (d) apply.

(a) Observation sample *(b) KDE result* *(c) Sampling grid* *(d) Sampled KDE*

Figure 2. Sampling procedure to transform point observations to a grid.

With a fixed sampling grid, this procedure results in a matrix of spatial observations for each data source, where each matrix element has a fixed correspondence to a specific geographic location. This procedure ensures that the spatial dependency between proximal observations is maintained and assigned a weight. Essentially, the preprocessing generates a consistent set of spatial observations among all data sources that can consequently be employed in PLS and regression analysis.

4.2 Identification of Latent Components using PLS

The crime data at hand refers to police records on 13 different crime types in San Francisco, each of which is probably individually depending on environmental factors. Our data sources census, map, and social media sum up to a total of 16 variables that reflect the environment in its various facets. These variables are shown in Figure 3, according to the category they fall into.

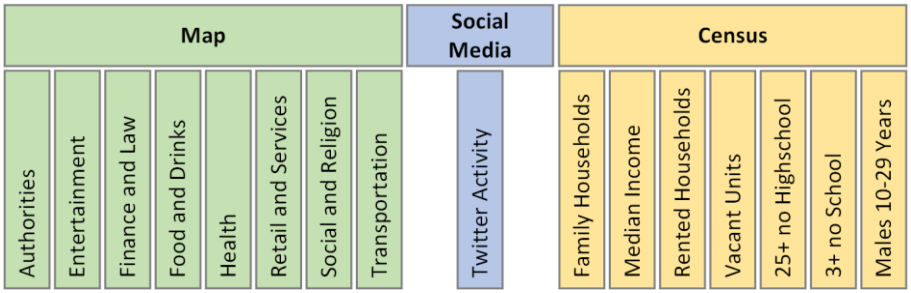


Figure 3. Original input variables by data source.

With these variables as a starting point, we seek for a deeper exploration of relations among the environmental measures. For this purpose, we carry out a series of PLS applications using each available crime type as dependent variable and our input variables (cf. Figure 3) as regressors. Since we follow an exploratory approach in seeking to find clusters among input variables and to explain their respective logical background, we start our PLS analyses with as many latent structures as possible. Consequently, we request PLS to estimate 16 latent variables to be sure that no information is lost during this procedure.

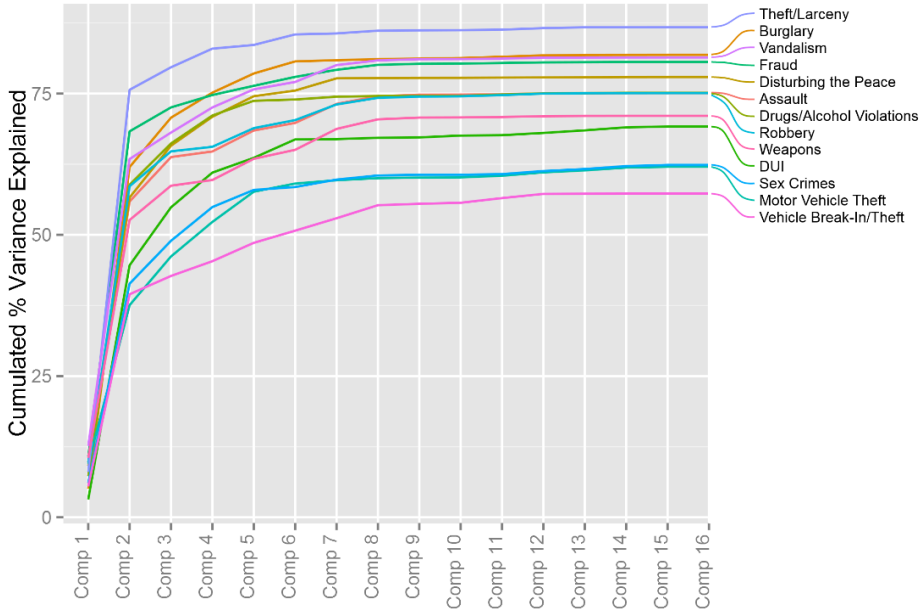


Figure 4. Cumulated variance explained for all crime types.

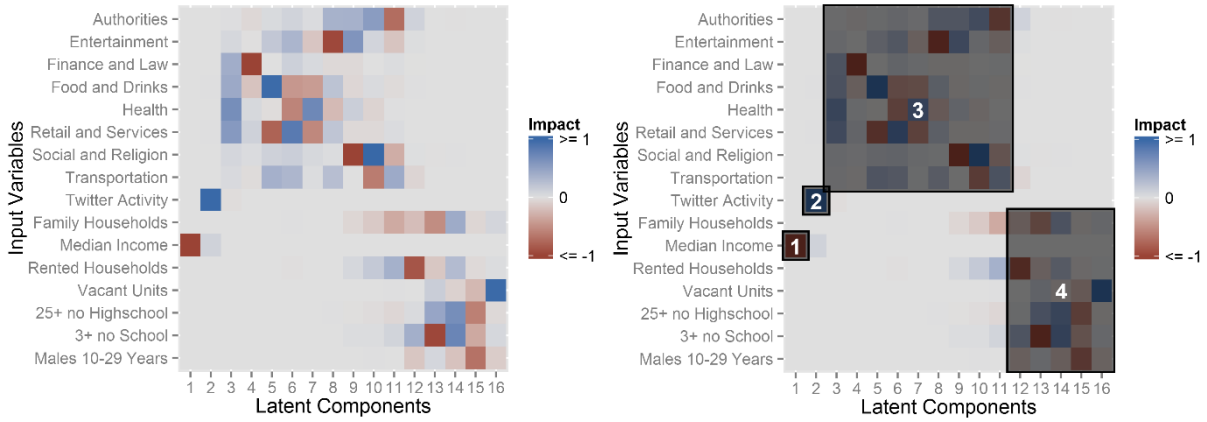
When applying the PLS methodology to a set of variables and requiring to estimate the same number of latent structures in the end, PLS essentially reconfigures the input measures. It efficiently unveils

combinations among the variables that maximize the explained variance for each latent component step by step. Figure 4 provides visual evidence of how the addition of latent components improves explained variable for each crime type. While the overall model cannot increase the explanatory value for obvious reasons, the order and combination of latent structures results in step-wise increase of explained variance. As a side note, PLS may estimate diverging latent components for different crime types; the actual linear combinations of input variables are not necessarily the same across crimes. However, we can see a similar structure across all crime types; the first two components are most valuable in terms of explained variance. The values steadily decrease over the upcoming components and fall below 1 in component 9 at the latest. This indicates that whatever is measured in the latter components will doubtlessly not have a remarkable impact on spatially explaining the emergence of criminal activity.

4.3 Logical Structure among Latent Components

In general, the latent structures resulting from PLS are not easily interpreted, since their combination is based on maximizing the step explanatory power step by step. Obviously, PLS is unable to optimize for human interpretability. Thus, we cannot solely rely on PLS' magic in reducing the amount of variables by linearly combining them into fewer new structures. One would usually visualize the residuals of a global model—i.e. by using a dendrogram—in order to identify the required number of latent components and employ PLS subsequently. In deep contrast, we are required to estimate the maximum possible number of latent structures to maintain interpretability and do the actual sense-making by ourselves. This approach is in line with Zheng and Pavlou (2010), who propose to manually prune the components identified by PLS into disjoint sets for further investigation. Interpretability is an aspect that quickly moves out of focus once automated estimations are applied.

The loadings identified by PLS for burglary crimes are depicted in Figure 5. Interestingly, the pattern of loadings—as shown in panel (a)—is very similar among all crime types, regarding both sign and value of each variable's impact on the latent components. Reading the figure column-wise, we can spot two almost singular components for median income and twitter activity. Ranging from component 3 to 11, the loadings are essentially depending on points of interest data, whereas components 12-16 mainly reflect census data. Panel (b) highlights these four clusters of dependency. We can already detect one of the clusters in the latter components we know to be less valuable for explanation due to their low explained variance.



(a) Latent components for burglary crimes *(b) Identified logical clusters among components*

Figure 5. PLS components for burglary crimes and their logical clustering.

For ease of understanding, we assign names to the four identified clusters. Since they are distinct in what input variables they depend on, naming is based on the logical commonalities in terms of input measures.

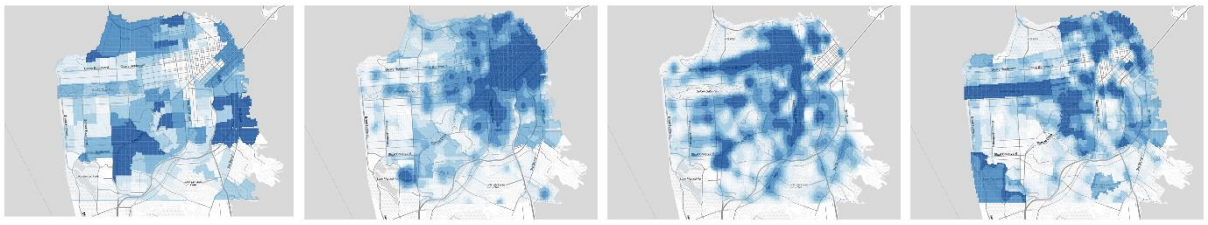
(1) **Prosperity.** The singular cluster mainly depends on the median income and thus reflects the wealth of residents in a specific area. This measure can possibly also refer to the potential value of theft-related crimes in the respective areas.

(2) **Presence.** Also the second cluster is singular and mainly depends on the Twitter activity. We can employ it to reflect the public presence, since each Twitter message is an indication that there has been a person in a specific location at a certain time. The aggregate of all these Twitter messages thus can mirror the density of people in an area.

(3) **Opportunity.** All latent components within this cluster represent points of interest and various combinations thereof. Hence, they reflect the physical environment. Specific characteristics of this environment may promote or suppress certain crime types, which can therefore be interpreted as the given environmental opportunity.

(4) **Readiness.** The fourth cluster is solely based on census data. As such, it represents the demographical and structural background of residents. Since this cluster does not cover the median income, it can be seen as representing the readiness of residents.

For further investigation, we combine the identified latent components into new variables according to their logical clusters outlined above. Their respective spatial distribution is outlined in Figure 6. We are aware of the fact, that aggregating latent components by their clusters may lead to a weaker model on the whole. However, the aggregation procedure is a tradeoff between maximizing explained variance and maintaining interpretability. Summing up the PLS loadings for each cluster makes us lose some fraction of explanatory power, since the explained variance is automatically averaged cluster-wise at the same time. While this may lead to slightly weaker results, we maintain a high degree of interpretability, which in our sense is superior.



Prosperity *Presence* *Opportunity* *Readiness*

Figure 6. PLS components for burglary crimes and their logical clustering.

However, even though we are able to assign a meaningful name to each of the clusters, the sense-making may still be ambiguous. We therefore have to clarify a few aspects concerning each of the clusters. First of all, prosperity majorly reflects the median income of residents. It does not contain information on how easily individual properties may be entered, or to what extent security efforts are taken. Altogether, prosperity may render as most relevant when explaining crimes against the property and belongings of individuals in general. Secondly, presence refers to the general public presence and does not measure public awareness in any quantifiable way. Specifically, this means that people who tweet do not necessarily represent those—if any—who have noticed a crime. The Twitter activity is mostly unrelated to specific criminal incidents. Thirdly, opportunity does only reflect the general (i.e. structural, physical) environmental conditions. It does not include any information on individual situational opportunities where a crime—for example theft—may seem promising and innocuous to an offender. Opportunity in the terminology of this research work refers to the overall possibilities that may emerge from different combinations of points of interest and their respective densities in an area. Fourthly, readiness is quantified by census measures on education, age, family structure, household type, and gender. It is measured in the areas where crimes are committed, not where the potential issuer grew up or came from. As such, it may rather be representative for those crimes where delinquents do not navigate to a specific, dedicated location for purposeful criminal activity. Even though median income is excluded, readiness could also reflect how promising a specific region is in terms of crime against body or property of people. Finally, nothing can represent the spatial characteristics of crime committed in the heat of the moment.

We will further investigate on these thoughts and aim to clarify interpretational ambiguity in the upcoming chapters. Since our research target covers explanation and prediction of criminal activity in a well-defined urban region, the upcoming analyses can be classified as meso-level. The clusters prosperity, presence, and readiness would rather suggest an individual level of analysis, but as they are employed in an aggregated form, we lose the references to single citizens, families, or households. Along with the opportunity cluster, the level of analysis resides on mid-range.

5 Assessing Inner Dependencies

An open door may tempt a saint is a well-known proverb which states that a promising situation may even lead the honest into temptation. Transferred to the scenario of this research work, it indicates that presence and opportunity may have a much higher impact on criminal activity than prosperity and readiness have.

Thus, the four identified clusters are probably not equally valuable when explaining or predicting criminal activity. From early visual inspection of the plots given in Figure 4 and Figure 5, we expect clusters to be generally more valuable the more sinistral their components are. This rationale is driven by the nature of PLS—early identified latent components are those that probably have the highest impact. As a first step, we carry out step-wise OLS regressions for each crime type to estimate the explanatory power of each single latent structure *presence*, *prosperity*, *opportunity*, and *readiness* after clustering. In a second step, we assess the dependencies among our four new latent structures by application of PLS path modeling.

5.1 Step-wise OLS Regression

In order to test and compare the respective impact of the four latent structures on different crime types, we perform regular OLS regressions for each crime and sequentially add the regressors one by one. The results outlined in Table 1 present the coefficient estimates for burglaries along with the respective t-statistics and significance codes for each regressor (row) in each successive step (column). In each step, all estimates are significant on a 0.001 level and—as a correlation test shows—all regressors are uncorrelated with the respective remaining ones. All estimates are stable across the steps, they do not vary when new variables are added to the model. Regarding the R^2 value, we can identify presence as the driving regressor when added in Step 2. Opportunity further increases the explanatory power, whereas prosperity and readiness only show marginal effects in terms of R^2. Both information criteria—AIC and BIC—state that the best model is obtained in Step 4 when all latent components are involved.

As expected in the preceding section, we are actually losing some part of our model's power when clustering the initial 16 latent components. The regular OLS model including the non-clustered latent components results in an adjusted R^2 of 0.818 and its information criteria AIC and BIC are, respectively,

	Step 0	Step 1	Step 2	Step 3	Step 4
(Intercept)	8.3 (130.8)	20.4 (87.8)	7.8 (48.7)	2.4 (18.3)	4.1 (27.5)
Prosperity		9.5 E−5 (38.2)	9.5 E−5 (60.4)	9.5 E−5 (79.1)	9.5 E−5 (79.8)
Presence			0.01 (202.6)	0.01 (265.0)	0.01 (267.5)
Opportunity				0.35 (139.6)	0.35 (140.9)
Readiness					2.1 (22.6)
Adjusted R^2	0	0.05	0.62	0.778	0.782
AIC	317143	220433	195357	180635	180133
BIC	317160	220458	195390	180676	180182

Stated: OLS coefficients based on 19094 degrees of freedom, t-statistics in parentheses.
All coefficients significant at 0.001 level

Table 1. Step-wise regression using latent components for burglary crimes.

175140 and 175288. Even though this complete model is superior to our resulting best model (Step 4 in Table 1), it is not interpretable at all. However, to allow for deeper sense-making, we take this tradeoff and stick to our model of clustered latent structures including prosperity, presence, opportunity, and readiness. In dimensions of R^2 speaking, the loss we face is lower than what prosperity—a single-component-cluster—has effectuated.

5.2 PLS Path Modeling

The previous steps of analysis give evidence that the identified latent structures—and especially their clustering—contain valuable and interpretable information on the emergence of various crime types. Our four clusters represent the public presence, the wealth of residents, the physical environment, and the demographical and sociological background of inhabitants. So far, we have assessed the explanatory power of each component independently. In a next step, we employ PLS path modeling to further research the dependencies and relations among the respective latent clusters presence, prosperity, opportunity, and readiness. For this purpose, we first identify a representative inner model that defines the connections among these four variables. We then estimate the edge weights to quantify the respective dependencies.

The PLS path modeling technique requires an outer model and an inner model. The outer, formative model describes the combination of original input variables that form the four latent structures as given in the last chapter. The inner model is also formative and describes how the four latent structures depend on each other and how they respectively influence delinquencies. Based on the goodness of fit (GoF) measure, we have identified the following model as the one to fit the data best among all valid combinations. Opportunity and readiness are described to be endogenous for the inner model, even though they are originally driven by data from the outer model in a global context. Figure 7 delineates the inner model and the edge weights how they apply to burglary crimes. The structure of the inner model is the same for all observed crime. However, the specific path weights and their signs vary among different delinquencies. In general, we can observe a strong dependency between presence and opportunity and a remarkable impact of opportunity on delinquency. In the specific case of burglaries, we can observe prosperity to have a negative effect on readiness but a positive impact on delinquency. Both presence and opportunity have a positive impact on all of their successor nodes.

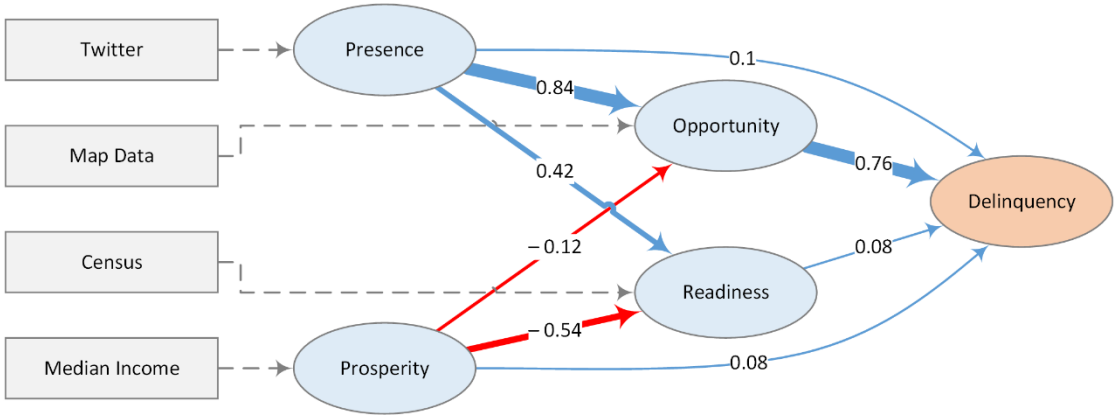
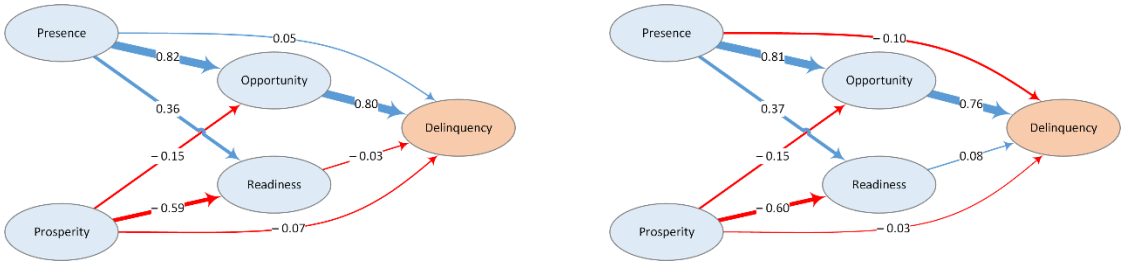


Figure 7. Formative inner model with path coefficients for burglary crimes.

As mentioned before, the specific path coefficients of the inner model and their respective signs may change depending on the crime type. Regarding assault crimes, outlined in Figure 8 (a), the paths from readiness to delinquency, and from prosperity to delinquency have negative signs in contrast to the burglary case. Essentially, we can infer that assaults are less likely to happen in areas of increased resident wealth and increased readiness. Panel (b) of Figure 8 depicts the inner model and path estimates for motor vehicle thefts. It states that especially public presence hinders delinquents from stealing cars in the close vicinity. Furthermore, cars are less likely to be stolen in areas of high prosperity. A possible explanation is that cars from residents with higher income—as such, probably more expensive cars—are rather not stolen when parked near their owners' homes. Such crime-specific dependencies are

observed for all analyzed crime types and they unveil relations among the latent nodes that are perceived to be reasonable. Based on these insights, we proceed to the next step and attempt to predict the emergence of different crimes.



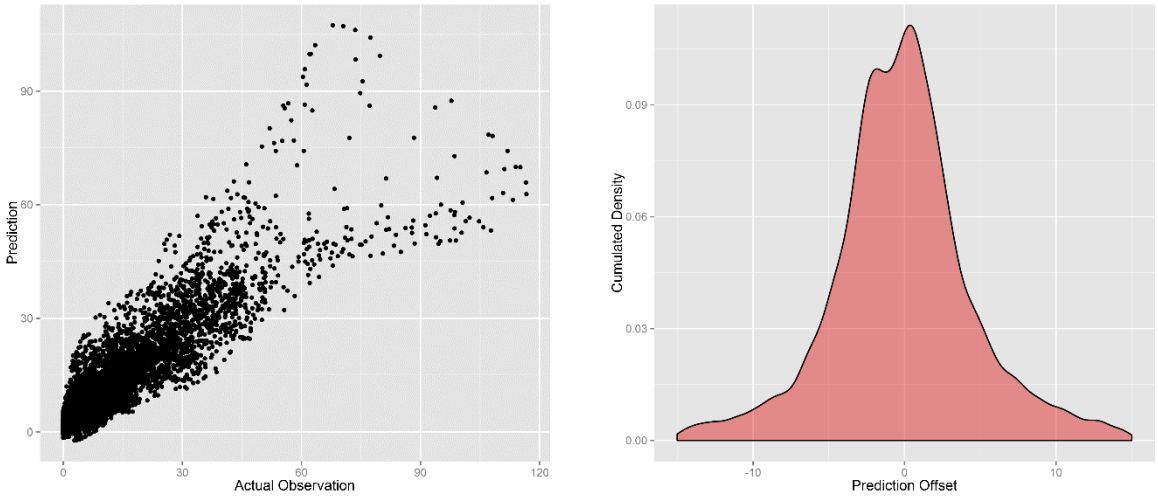
(a) Assault *(b) Motor Vehicle Theft*

Figure 8. Inner model evaluation of assault and motor vehicle theft

6 Crime Prediction using Latent Structures

As the final step of our analysis, we now attempt to predict crimes solely based on the four identified latent structures prosperity, presence, opportunity, and readiness. For our predictive approach, we split our data set into a training set and a test set at ratio 7:3. Since our data consists of spatial observations, we have to permute all observations prior to selecting 70 per cent for the training set. Otherwise, we would not dissolve the spatial correspondence and therefore would predict data at geographical spots without having used their vicinity for training. With training and test sets at hand, we perform a regular OLS regression using our four latent measures and linearly calculate the prediction for our test set.

Figure 9 shows two plots that delineate the prediction accuracy and offset density. In panel (a), the actual crime observations are compared to the calculated prediction. We can clearly see the linear dependency, which states that our prediction is quite accurate. At the same time, the contained heteroskedasticity is obvious, which means that our prediction gets increasingly imprecise where more crimes are likely to happen. The prediction deviation density is depicted in panel (b). It shows a symmetrical distribution, effectively stating that the prediction does not tend towards over- or underestimation. Overall, the mean squared error of 0.0209 means that we miss the perfect prediction by 0.145 crimes on an average.



(a) Prediction accuracy *(b) CDF for prediction deviation*

Figure 9. Inner model evaluation of assault and motor vehicle theft

The obvious heteroscedasticity of the prediction is confirmed by a Breusch-Pagan test. Consequently, the regression coefficients require for correction to remain interpretable. The OLS coefficients of the training regression using our four latent structures are outlined in Table 2 with their t-values already

corrected for heteroscedasticity. The correction procedure reduces the t-values, but all estimates remain significant on a 0.001 level and thus their explanatory and predictive power is maintained. All in all, predicting criminal activity from environmental characteristics works exceptionally well, independently from the certain type of delinquency.

	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	3.97 E+00	1.91 E–01	20.73	<2 E–16	***
Presence	1.02 E –02	1.32 E–04	77.04	<2 E–16	***
Prosperity	9.40 E –05	2.03 E–06	46.41	<2 E–16	***
Opportunity	3.52 E –01	4.80 E–03	73.34	<2 E–16	***
Readiness	2.07 E+00	1.40 E–01	14.85	<2 E–16	***
Adj. R^2	0.7815				
Coefficients based on 19094 degrees of freedom. Significance codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1					

Table 2. OLS regression results for burglary crimes.

7 Concluding Remarks

The regression results outlined in Table 5 unveil that the four latent structures have a different impact when examining spatial characteristics of crime. Following their t-values, presence and opportunity are most valuable. In terms of expressive power, prosperity is the follow-up and readiness is least valuable. In order to assess their respective explanatory potentials, we discuss their interplay and attempt to interpret the observed effects.

From the inner path model, we are aware that presence has a significantly high impact on opportunity in general. Thus, we can interpret the public presence to reflect one facet of opportunity. Nonetheless, it may have a varying impact depending on the certain crime type and in turn is better employed as a separate measure. Prosperity as a measure that reflects the wealth of residents may be a driving factor for crimes that are aimed against the property of others. Hence, its impact is highest when exploring the emergence of, for instance, burglary, vehicle break-in/theft, or motor vehicle theft. Opportunity is highly valuable for any kind of criminal activity, most likely because it reflects the physical structure of the city as no other measure can. It accounts for spatial distribution and density of points of interest and therefore at the same time reflects the natural structure of urban regions. Finally, readiness represents the structural, educational, and financial situation of inhabitants by region. As such, it has no answer to questions concerning the delinquent if he does not reside where the crime was committed and recorded. We can suspect that, with some exceptions, most criminals do not commit purposeful crime in very close vicinity of their homes. This assumption is in line with Canter and Larkin (1993). The authors distinguish between commuter and marauder behavior using sex crimes as an example. Following a commuter behavior, delinquents select a promising area for their crimes and return regularly for further activity. In contrast, the marauder behavior means to search for targets around—but not too close to—the own location. Depending on the crime type, we may observe one or the other, essentially rendering readiness useful or not. Being subject to such uncertainty and at the same time being employed in many studies nonetheless, we have to ask whether readiness is valuable for spatial explanation or prediction of criminal activity at all. As a final addition, nothing can reflect crimes that are committed in the heat of the moment.

In future research, we plan to compare our findings to insights from other cities across the world. Furthermore, we want to test the applicability of our proposed methodology on spatial observations other than crimes. Many spatial observations would be appropriate that include human activity and behavioral patterns in urban regions. In case it turns out to be valuable for a broader domain, it paves the way for deeper analyses and valuable insights for many businesses, local authorities, and governments. Moreover, IS research in general can provide valuable insights in this field resulting from combinations of novel big data approaches and behavioral as well as sociological considerations.

References

Andresen, M. A. 2006. "Crime Measures and the Spatial Analysis of Criminal Activity," *British Journal of Criminology* (46:2), pp. 258–285.

Antonakis, J., Bendahan, S., Jacquart, P., and Lalive, R. 2014. "Causality and Endogeneity: Problems and Solutions," in *The Oxford handbook of leadership and organizations*, D. V. Day (ed.), Oxford, New York, NY: Oxford University Press, pp. 93–117.

Bachner, J. 2013. *Predictive Policing: Preventing Crime with Data and Analytics*: IBM Center for The Business of Government .

Bendler, J., Ratku, A., and Neumann, D. 2014a. "Crime Mapping through Geo-Spatial Social Media Activity," in *35th International Conference on Information Systems (ICIS 2014)*.

Bendler, J., Wagner, S., Brandt, T., and Neumann, D. 2014b. "Taming Uncertainty in Big Data," *Business & Information Systems Engineering* (6:5), pp. 279–288.

Bertot, J. C., Jaeger, P. T., and Grimes, J. M. 2010. "Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies," *Government Information Quarterly* (27:3), pp. 264–271.

Block, R. 1979. "Community, Environment, and Violent Crime," *Criminology* (17:1), pp. 46–57.

Bookstein, F. L. 1980. "Data Analysis by Partial Least Squares," in *Evaluation of econometric models*, J. Kmenta and J. B. Ramsey (eds.), New York: Academic Press, pp. 75–90.

Bowers, K. 1999. "Exploring links between crime and disadvantage in north-west England: an analysis using geographical information systems," *International Journal of Geographical Information Science* (13:2), pp. 159–184.

Bowes, D. R. 2007. "A Two-Stage Model of the Simultaneous Relationship Between Retail Development and Crime," *Economic Development Quarterly* (21:1), pp. 79–90.

Brantingham, P. L., and Brantingham, P. J. 1993. "Nodes, paths and edges: Considerations on the complexity of crime and the physical environment," *Journal of Environmental Psychology* (13:1), pp. 3–28.

Canter, D., and Larkin, P. 1993. "The environmental range of serial rapists," *Journal of Environmental Psychology* (13:1), pp. 63–69.

Chainey, S., Tompson, L., and Uhlig, S. 2008. "The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime," *Security Journal* (21:1-2), pp. 4–28.

Chang, R. M., Kauffman, R. J., and Kwon, Y. 2014. "Understanding the paradigm shift to computational social science in the presence of big data," *Decision Support Systems* (63), pp. 67–80.

Cohen, J. 1941. "The Geography of Crime," *Annals of the American Academy of Political and Social Science* (217), pp. 29–37.

Curman, A. S. N., Andresen, M. A., and Brantingham, P. J. 2015. "Crime and Place: A Longitudinal Examination of Street Segment Patterns in Vancouver, BC," *Journal of Quantitative Criminology* (31:1), pp. 127–147.

Day, D. V. (ed.) 2014. *The Oxford handbook of leadership and organizations*, Oxford, New York, NY: Oxford University Press.

Dijkstra, T. K., and Henseler, J. 2015. "Consistent Partial Least Squares Path Modeling," *MIS Quarterly*.

Dijkstra, T. K. 2010. "Latent Variables and Indices: Herman Wold's Basic Design and Partial Least Squares," in *Handbook of Partial Least Squares*, V. Esposito Vinzi, W. W. Chin, J. Henseler and H. Wang (eds.), Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 23–46.

Effing, R., van Hillegersberg, J., and Huibers, T. 2011. "Social Media and Political Participation: Are Facebook, Twitter and YouTube Democratizing Our Political Systems?" in *Electronic Participation*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, E. Tambouris, A. Macintosh and H. de Bruijn (eds.), Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 25–35.

Gerber, M. S. 2014. "Predicting crime using Twitter and kernel density estimation," *Decision Support Systems* (61), pp. 115–125.

Grubesic, T. H., and Mack, E. A. 2008. "Spatio-Temporal Interaction of Urban Crime," *Journal of Quantitative Criminology* (24:3), pp. 285–306.

Henseler, J., Ringle, C. M., and Sinkovics, R. R. 2009. "The use of partial least squares path modeling in international marketing," in Sinkovics, R. R. and Ghauri, P. N. (eds.) *New Challenges to International Marketing* (*Advances in International Marketing, Volume 20)*, Bingley: Emerald Group Publishing.

Henseler, J., and Sarstedt, M. 2013. "Goodness-of-fit indices for partial least squares path modeling," *Computational Statistics* (28:2), pp. 565–580.

Hulland, J., Ryan, M. J., and Rayner, R. K. 2010. "Modeling Customer Satisfaction: A Comparative Performance Evaluation of Covariance Structure Analysis Versus Partial Least Squares," in *Handbook of Partial Least Squares*, V. Esposito Vinzi, W. W. Chin, J. Henseler and H. Wang (eds.), Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 307–325.

Irvin, R. A., and Stansbury, J. 2004. "Citizen Participation in Decision Making: Is It Worth the Effort?" *Public Administration Review* (64:1), pp. 55–65.

Krivo, L. J., and Peterson, R. D. 1996. "Extremely Disadvantaged Neighborhoods and Urban Crime," *Social Forces* (75:2), pp. 619–648.

Lee, K. C., Lee, D. S., Seo, Y. W., and Jo, N. Y. 2011. "Antecedents of Team Creativity and the Mediating Effect of Knowledge Sharing: Bayesian Network Approach to PLS Modeling as an Ancillary Role," in *Intelligent Information and Database Systems*, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, N. T. Nguyen, C.-G. Kim and A. Janiak (eds.), Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 545–555.

McIntosh, C. N., Edwards, J. R., and Antonakis, J. 2014. "Reflections on Partial Least Squares Path Modeling," *Organizational Research Methods* (17:2), pp. 210–251.

Messatfa, H., Ryes, L., and Schroeck, M. 2011. "The power of analytics for public sector: Building analytics competency to accelerate outcomes," IBM Institute for Business Value (IBM Global Business Services).

Murray, A. T. 2001. "Exploratory Spatial Data Analysis Techniques for Examining Urban Crime: Implications for Evaluating Treatment," *British Journal of Criminology* (41:2), pp. 309–329.

Nelson, A., Bromley, R., and Thomas, C. 2001. "Identifying micro-spatial and temporal patterns of violent crime and disorder in the British city centre," *Applied Geography* (21:3), pp. 249–274.

Patterson, E. B. 1991. "Poverty, Income Inequality, and Community Crime Rates," *Criminology* (29:4), pp. 755–776.

Ratcliffe, J. 2004. "The Hotspot Matrix: A Framework for the Spatio-Temporal Targeting of Crime Reduction," *Police Practice and Research* (5:1), pp. 5–23.

Ratcliffe, J. 2010. "Crime Mapping: Spatial and Temporal Challenges," in Handbook of Quantitative Criminology, A. R. Piquero and D. Weisburd (eds.), New York, NY: Springer New York, pp. 5–24.

Roncek, D. W. 1981. "Dangerous Places: Crime and Residential Environment," *Social Forces* (60:1), pp. 74–96.

Tobler, W. R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region," *Economic Geography* (46), p. 234.

Traunmueller, M., Quattrone, G., and Capra, L. 2014. "Mining Mobile Phone Data to Investigate Urban Crime Theories at Scale," in *Social Informatics*, L. M. Aiello and D. McFarland (eds.), Cham: Springer International Publishing, pp. 396–411.

Vinzi, V. E., Trinchera, L., and Amato, S. 2010. "PLS Path Modeling: From Foundations to Recent Developments and Open Issues for Model Assessment and Improvement," in *Handbook of Partial Least Squares*, V. Esposito Vinzi, W. W. Chin, J. Henseler and H. Wang (eds.), Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 47–82.

Wang, X., Gerber, M. S., and Brown, D. E. 2012. "Automatic Crime Prediction Using Events Extracted from Twitter Posts," in *Social Computing*, Behavioral - Cultural Modeling and Prediction, D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, M. Naor, O. Nierstrasz, C. Pandu Rangan, B. Steffen, M. Sudan, D. Terzopoulos, D. Tygar, M. Y. Vardi, G. Weikum, S. J.

Yang, A. M. Greenberg and M. Endsley (eds.), Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 231–238.

Wold, S., Eriksson, L., and Kettaneh, N. 2010. "PLS in Data Mining and Data Integration," in *Handbook of Partial Least Squares*, V. Esposito Vinzi, W. W. Chin, J. Henseler and H. Wang (eds.), Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 327–357.

Zheng, Z., and Pavlou, P. A. 2010. "Research Note —Toward a Causal Interpretation from Observational Data: A New Bayesian Networks Method for Structural Models with Latent Variables," *Information Systems Research* (21:2), pp. 365–391.