Dynamic Network Model for Smart City Data-Loss Resilience Case Study: City-to-City Network for Crime Analytics

Olivera Kotevska, A. Gilad Kusne, Daniel V Samarov, Ahmed Lbath, Abdella Battou

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCED DATA ANALYTICS FOR LARGE-SCALE COMPLEX DATA ENVIRONMENTS

Received July 5, 2017, accepted September 11, 2017, date of publication October 12, 2017, date of current version October 25, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2757841

Dynamic Network Model for Smart City Data-Loss Resilience Case Study: City-to-City Network for Crime Analytics

OLIVERA KOTEVSKA[1], (Student Member, IEEE), A. GILAD KUSNE[1], DANIEL V. SAMAROV[1], AHMED LBATH[2], (Member, IEEE), AND ABDELLA BATTOU[1]

[1]National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
[2]University of Grenoble Alpes, 38400 Grenoble, France

Corresponding author: Olivera Kotevska (olivera.kotevska@nist.gov)

This work was supported by the Information Technology Laboratory, National Institute of Standards and Technology.

ABSTRACT Today's cities generate tremendous amounts of data, thanks to a boom in affordable smart devices and sensors. The resulting big data creates opportunities to develop diverse sets of context-aware services and systems, ensuring smart city services are optimized to the dynamic city environment. Critical resources in these smart cities will be more rapidly deployed to regions in need, and those regions predicted to have an imminent or prospective need. For example, crime data analytics may be used to optimize the distribution of police, medical, and emergency services. However, as smart city services become dependent on data, they also become susceptible to disruptions in data streams, such as data loss due to signal quality reduction or due to power loss during data collection. This paper presents a dynamic network model for improving service resilience to data loss. The network model identifies statistically significant shared temporal trends across multivariate spatiotemporal data streams and utilizes these trends to improve data prediction performance in the case of data loss. Dynamics also allow the system to respond to changes in the data streams such as the loss or addition of new information flows. The network model is demonstrated by city-based crime rates reported in Montgomery County, MD, USA. A resilient network is developed utilizing shared temporal trends between cities to provide improved crime rate prediction and robustness to data loss, compared with the use of single city-based auto-regression. A maximum improvement in performance of 7.8 % for Silver Spring is found and an average improvement of 5.6 % among cities with high crime rates. The model also correctly identifies all the optimal network connections, according to prediction error minimization. City-to-city distance is designated as a predictor of shared temporal trends in crime and weather is shown to be a strong predictor of crime in Montgomery County.

INDEX TERMS Adaptive algorithms, geospatial analysis, predictive models, statistical learning.

I. INTRODUCTION

Smart city design seeks to optimize city services, by improving the resident experience and reducing waste, through intelligent use of citywide data. Smart city services are expected to respond appropriately to changing conditions, requiring regular data updates on the status of citywide properties, such as weather, road conditions, and communicable disease case numbers. Consequently, optimal deployment of critical resources will depend on data streams. For instance, in the event of a disease epidemic, current medical statistics will be used to ensure that ambulances, drugs, and vaccine resources are intelligently distributed to the worst hit neighborhoods and those predicted to be at high risk. In the case of a powerful storm that disrupts traffic, traffic data will be used to distribute police resources for traffic guidance based on current and predicted traffic patterns. In the case of changing crime rates throughout the city, crime statistics and predictions will be used to ensure that police, medical, and emergency resources are intelligently distributed to reduce response time.

As city services become more dependent on smart city data streams, the services also become more susceptible to disruptions in the data streams. Such disruptions can affect critical services, for instance, by increasing ambulance response time. Interruptions in data streams and the resulting loss of data can occur for any number of reasons, including anomalous signal to noise reduction, power loss during data collection, or data loss on a network level, either benign or malicious. Additionally, due to the spatial and temporal dependence of smart city data streams, with data collected periodically by distributed sensors or local human-based reporting, spatiotemporal events can impact data service. For example, a storm can knock out neighborhood-wide communications, interrupting data collection as the storm travels from one location to the next. For these reasons, smart city services require a level of resilience to data stream disruption based on application domain and data rate frequency.

A diverse set of techniques exists for establishing resilience to data stream disruptions, and the resulting data loss, at different network layers [16]. The work presented here focuses on resilience techniques at the application level. For example, if data loss occurs in one field or element of multi-field data, the entire entry can be removed, or a marked as missing. In particular, when knowledge of the missing data is required, many methods exist for estimating the missing data including the use of mean or median, extrapolation from past data, or the use of model-based methods such as maximum likelihood or multiple imputations [8]. To improve smart city resilience to data loss, such a scheme must be implemented. As data disruptions occur and data loss is identified, the lost data is estimated with minimal prediction error to reduce the impact of the data loss on dependent services. In this work, we propose an application-layer algorithm that can be used to ensure robustness across regions of different scale, from smart communities to smart counties, and establish both inter- and intra-smart city networks. Inter-smart city networks are of interest for county-wide occurrences such as the spread of epidemics, while intra-smart city network can provide resilience for the scenarios as traffic management. The algorithm identifies a dynamic data sharing network between independent, smart cities or generalized smart community to ensure minimal data estimation error for each smart city, with each smart city, assumed to be associated with a single multivariate data stream.

The proposed model incorporates three key features: (1) the ability to handle multivariate time series data streams and capitalize on temporal trends in past data to improve estimation accuracy, (2) identification and use of mutual information, in the form of statistically significant temporal trends shared between data streams, to improve estimation accuracy, and (3) network dynamics to respond to changing data and city conditions. The ability to handle multivariate time-series/spatiotemporal data streams is essential for many smart city applications as data is often collected for multiple, distributed locations and periodically over the same time range. For example, a set of data streams may pertain to hourly temperature data, with one data stream for each neighborhood. Furthermore, multivariate time-series techniques allow data analysts to identify and track temporal trends within a data stream to improve data prediction, increasing robustness in the presence of potential data loss. Additionally, shared information between data streams can be utilized to improve data prediction and increase robustness even further. For example, if a neighborhood weather sensor network experiences data loss at one node, data from a neighboring sensor can be used to reinforce estimation of the lost data. Finally, a practical, smart city resilience model must also be dynamic – capable of self-adaptation to changes in the city environment to maintain service. In the case of a crime incident data network, the network should be resilient to the closing of a reporting office in one location, or the opening of a new office. Similarity, dynamics is important in the case of travel data for travel safety, as a path that was safe yesterday may not be safe today. The three features are incorporated into the proposed resilient data-sharing network, ensuring optimal use of data in the face of potential disruptions and a dynamic environment.

The data sharing network is represented by a graph, with each smart city and its respective data stream indicated by one node, and directed edges are indicating data sharing connections between the smart cities. Data sharing is described by a set of vector-autoregression (VAR) [21] time series models which provide multivariate temporal analysis while capitalizing on shared temporal trends. The set of potential VAR models for each smart city is reduced by an automatic analysis of shared temporal trends between cities using Granger causality [9]. Both VAR and the Granger causality methods are popular methods in econometrics, used for predictive analytics in the stock market and exchange rate volatility [19]. The VAR model selected for each city utilizes data from cities with mutual information as indicated by Granger causality while also providing minimum data estimation error. The computational issue of model selection is further reduced by using Multidimensional Data Scaling (MDS) [8], a common data dimensional reduction technique, to investigate potential underlying environment variables that are predictive of model performance. These predictive environmental variables may then be used to whittle down the space of possible models. Dynamics in the data sharing network is achieved through a regular update of network connections performed by iterative analysis of data sharing efficacy.

The system is demonstrated on the application domain of crime data analytics for Montgomery County (MC), Maryland USA. A cross-county resilience network of city-to-city data sharing is identified and demonstrated to provide improved crime statistics prediction and data-loss resilience, as compared to analysis using city-based autoregression. The identified resilience network can be utilized to optimize medical, police, and emergency services as well as suggest policy changes enhance public safety and health [12]. The experimental data is comprised of the number of police-reported incidents, organized by city, throughout MC

between 01/01/2014 and 06/26/2016,[1] constituting a spatiotemporal multivariate time series with each data stream reporting one variable – the number of daily city-wide crime incidents. The resilience network capitalizes on underlying temporal trends within a city and shared temporal trends between cities to improve crime prediction, thus mitigating the effects of data loss and day-to-day crime rate variability. An investigation was also conducted to discover potential underlying demographic and topology parameters that may explain the evaluated network graph.

This paper is structured as follows. In section 2, we describe prior related work. Section 3 describes the problem statement followed by the proposed novel approach and a description of evaluation metrics. Section 4 examines potential network models for the Montgomery County crime dataset as well as their performance and analysis, followed by a discussion of the proposed network. Finally, Section 5 discusses the conclusions and a description of future work.

II. BACKGROUND

Prior work in smart city resilience models includes a variety of techniques and application areas, combining subsets of the three key data analysis features described in the introduction. We look at different statistical methods for resilience. For example, Anava *et al.* [4] only uses its own past predictors to overcome the missing data. Dogra and Kobti [7] incorporated dynamics and data sharing into a complex modeling system by utilizing an agent-based approach capable of evolving in response to a changing environment. As individual agents detect changes, those changes are shared with the other agents. Dynamics for fault tolerance were introduced in [5], which utilizes a Byzantine fault tolerance method to create hardware resilience to malicious attacks. These methods provide system resilience and adaptability to environmental changes but do not utilize temporal trends from past data to improve future resilience. Aman *et al.* [3] deployed dynamics, a temporal trend sensitive system using a combination of autoregressive integrated moving average (ARIMA) time series models to respond to the dynamics of energy demand. Similarly, [2] combines dynamics and time series based regression via a decision tree to improve prediction of complex events. Their algorithm identifies current model prediction error and dynamically determines to increase or decrease the time series training window accordingly.

Methods that leverage mutual information between data streams to improve estimation accuracy and robustness have also been used for smart city data-loss resilience, in particular for the application domains of weather and transportation. For example, the work in [20] combines support vector machine regression and a data network between neighboring weather sensors to interpolate missing spatiotemporal weather data from one sensor using neighboring sensors. However, this method assumes a static network, where network connections do not change over time, and where connections are defined by sensor proximity and similarity. A network of weather data streams is also addressed by Derguech *et al.* [6] where network dynamics is allowed with potential changes in data sharing connections. A greedy algorithm selects data streams with the most recent data and rejects data streams identified to provide poor prediction performance. Pearson's product moment correlation function is used to determine and evaluate data streams for shared temporal trends dynamically, and these learned relationships are then utilized across a set of regression techniques to ensure system resilience to data faults. Pravilovic *et al.* [17] also utilize correlation in their geo-sensor data resilience networks, where the temporal and spatial correlation between data streams are identified and used to establish a spatial-based cluster of data streams. A stationary correlation is assumed, and a static data sharing network is formed. However, while correlation analysis can provide useful information on shared temporal trends, it does not give an indication of the statistical significance of these relationships. Additionally, highly useful shared trends between data sources separated by significant distances may be missed in such spatial correlation based techniques.

The particular application domain of crime prediction and analysis has greatly benefited from the development of such smart city data analysis techniques, as well as an increase in data collection. Data analysis techniques are used to identify spatiotemporal patterns in crimes incidents and develop crime prediction models [12], [15]. For example, analyses of shared trends across locations are used to discover spatial patterns in crime incidents [13], and time series analysis techniques are used to find relevant and meaningful temporal patterns [13]. Spatiotemporal crime trend analysis, which studies the dynamic interplay of location-dependent and time-dependent aspects of crime, utilizes a wide variety of techniques including pattern mining, association rule mining, and combinations of the previously mentioned methods [13]. Many methods rely on the use of multivariate time series or relationship analysis to improve crime prediction. For example, Gunderson et al. demonstrated in [10] a time series multi-agent model to predict areas in which future criminal incidents are likely to happen and use both physical and cyber-criminal activity data. Liao *et al.* [14] utilize Bayesian inference to create a geographical map to show potential crime factors per area. This weighted geographical profile provides probability estimation for the next crime hot spots and likely locations for future crime incidents. Such results can be used to improve police resource deployment. Gerber *et al.* applied in [9] a Latent Dirichlet Allocation semantic analysis to capitalize on the relationship between criminal activity and Twitter crime discussions to identify crime-predictive Twitter discussion topics. Ranson *et al.* [18] used linear regression to investigate the relationship between weather factors and crime data and identified a strong relationship between climate and crime incident number and type. These crime statistics studies have used a subset of the

[1] https://data.montgomerycountymd.gov/

three key features of the proposed algorithm and may benefit from expanding the data analysis technique to use all three features.

III. PROBLEM DESCRIPTION AND PROPOSED SOLUTIONS

A. PROBLEM STATEMENT AND NOTATION

Smart city service optimization depends on linking services to the dynamic state of the city, measured by time series data collected across the city. The challenge is to identify a flexible mechanism to ensure optimal, resilient, smart city services in the presence of potential data loss within a data stream, reduction in data stream quality, the loss of an entire data stream, or the addition of a new data stream. The system should also capitalize on shared information between data streams to ensure optimal performance.

For this work, sets of multivariate, spatiotemporal smart city data streams such as the status of multiple traffic lights, the number of locally available vaccine units, and neighborhood air quality are represented by $\boldsymbol{Y} = \left[\boldsymbol{y}^{k=1}, \boldsymbol{y}^{k=2}, \ldots, \boldsymbol{y}^{k=N}\right]$, where the superscript $k \in \{1, \ldots, N\}$ provides the data stream index for a set of N data streams. For spatial data, each index k corresponds to a location. Individual data streams are represented by a time-series vector $\boldsymbol{y}^l = \left[y^l_{t=0}, y^l_{t=-1}, \ldots, y^l_{t=-v}\right]$ with the subscript t providing the time series sample index, beginning at the time of interest to be predicted $t = 0$ and extending to v periods in the past $t = -v$. An individual data stream $\boldsymbol{y}^l$ thus has dimensions $\mathbb{R}^p$ and the set of N data streams $\boldsymbol{Y}$ has dimensions $\mathbb{R}^{N \times p}$. A snapshot of the state across all streams at time t is given by $\boldsymbol{y}_t = \left[y^1_t, y^2_t, \ldots, y^N_t\right]$. For this work, we assume that all data streams are simultaneously sampled at regular time intervals. Data loss in a data stream is indicated by the absence of data at a time, $y^l_t = \emptyset$.

The estimate for data value y^l_0 is represented by $\hat{y}^l_0$. When using the set of data streams $\left\{\boldsymbol{y}^l, \boldsymbol{y}^m, \boldsymbol{y}^n\right\}$ to evaluate $\hat{y}^l_0$, the functional relationship is represented by $\hat{y}^l_0 = f\left(\boldsymbol{y}^l, \boldsymbol{y}^m, \boldsymbol{y}^n\right)$. For this work, it is always assumed that $\hat{y}^l_0$ is estimated using past data $t < 0$ from the independent data streams. For example, the set of three data streams can be the daily crime incident numbers for two cities indexed by l and m, and n is the daily regional temperature. Here we use past data from these three data streams to estimate $\hat{y}^l_0$, the crime incident number on day $t = 0$ for city l. The functional relationship between data streams can be represented using a directed graph, with data streams indicated by nodes, and directed edges connecting nodes that represent independent data streams to nodes representing dependent data stream. As both AR and VAR models require the self-dependent functional relationship $\hat{y}^l_t = f\left(\boldsymbol{y}^l\right)$ for any t, the edge pointing from $\boldsymbol{y}^l$ back to $\boldsymbol{y}^l$ is assumed and not indicated in the presented figures. The set of all concurrent resilient models composes the network graph, $G = \left\{f^{i=1}, f^{i=2}, \ldots, f^{i=K}\right\}$ with i indexing the set of all K functions.

B. PROPOSED SOLUTION

We propose a dynamic network-based model that provides improved and reinforced resilience to data loss. The VAR-based model utilizes past data from the data stream of interest as well as data from 'related' data streams that share temporal trends, to achieve optimal estimation accuracy. The model dynamically identifies the optimal set of data sources to reinforce the data-loss robustness of each data stream with Granger causality and MDS analysis. Model dynamics is achieved through recurrent updates, which identify the optimal network connections for each data stream to maintain optimal estimation accuracy.

An example is shown in Figure 1. Three data streams are presented with their values indicated for times t_{-4} through t_5. At time t_0, the data stream of interest $\boldsymbol{y}^1$ experiences data loss. (At this time the data for times t_1 through t_5 have yet to be collected.) Resilience in the data stream can be established by estimating the lost data using autoregression (AR) – extrapolating the value of y^1_0 from past data. Alternatively, if either available data stream $\boldsymbol{y}^2$ or $\boldsymbol{y}^3$ shows similar trends to data stream $\boldsymbol{y}^1$, information from that stream can be used to improve the estimate of y^1_0 using VAR. During the period of $t = \{-4, \ldots, 5\}$, there are four potential resilience models which can be evaluated for their utility in reinforcing estimation of y^1_t:

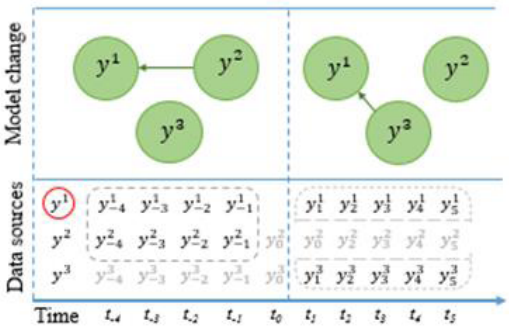

FIGURE 1. An illustration of three event-based data sources [y_1, y_2, y_3] and dynamic model adaptation over time [$t_{-4}, \ldots, t_5$] depending on data stream changes.

1) AR using only data stream $\boldsymbol{y}^1$: $\hat{y}^1_0 = f\left(\boldsymbol{y}^1\right)$.
2) VAR using data streams $\boldsymbol{y}^1$ and $\boldsymbol{y}^2$: $\hat{y}^1_0 = f\left(\boldsymbol{y}^1, \boldsymbol{y}^2\right)$.
3) VAR using data streams $\boldsymbol{y}^1$ and $\boldsymbol{y}^3$: $\hat{y}^1_0 = f\left(\boldsymbol{y}^1, \boldsymbol{y}^3\right)$.
4) VAR using all three data streams: $\hat{y}^1_0 = f\left(\boldsymbol{y}^1, \boldsymbol{y}^2, \boldsymbol{y}^3\right)$

Data streams two and three are first tested for shared trends with data stream 1 using the Granger causality test to determine the viability of models 2-4. The Granger causality test [9], as well as MDS, are described in section 3.2b. If shared trends are identified, the models are queried for their performance at each period, the models are ranked by performance, and the best performing model is selected for implementation. Performance is measured by computing estimated prediction error using time series cross-validation. If instead a supporting data stream is found not to provide utility, that data stream can be removed from the later analysis, reducing the

amount of data traffic required for the network. In this example, model 2 is determined to provide optimal performance for period $t = \{-4, \ldots, 0\}$ and once data has been collected for $t = \{1, \ldots, 5\}$, model 3 is found to provide the best performance for this period. Models 2 and 3 are graphically represented by a directed graph, with edges connecting the nodes representing y^2 or y^3 to the node representing y^1 (See Figure 1). If an issue is identified with the optimal resilience model, e.g., the supporting data stream stops reporting, then the next best performing model is chosen, and so on.

The set of all top performing, concurrent models for all data streams composes the resilience network. The network is represented by the resilience network graph – the collected graphical representation of all concurrent models. At user-determined intervals, the system is iteratively updated, re-evaluating the performance of each model to update model rankings and identify and implement the optimal model.

The resilience network method is diagramed in Figure 2, with each step explained below.

FIGURE 2. Overview of the proposed solutions – Data analytics framework with periodically based iterations. The right side represents the first two steps which constitute the data preprocessing (a) step: (a′) data wrangling and (a″) data normalization. Next step is data stream relationship analysis (b), used to identify the streams that share temporal trends and narrow down the hypothesis space of potential data sharing models for the network. Step (c) is to determine the list of best models for analysis based on minimum prediction error, and step (d) is dynamic respond on the best model selection and available resources.

The system begins with preprocessing the data streams, described in section 3a, followed by relationship analysis for sets of streams. Relationship analysis is performed to reduce the search space for optimal resilience models, as described in section 3b. Potential resilience models are then evaluated, the optimal models are selected, and the network is identified. These steps are iterated at user-determined intervals to maintain an updated, optimal resilience network. Qualitative analysis is also used to determine correlations between the network and any pertinent data relating the data streams. Through this qualitative analysis, additional information sources can be used to reduce the search space of potential resilience models and subsequently reduce computation time and cost.

1) DATA PREPROCESSING

Data preprocessing can involve many steps including data wrangling, and feature vector normalization. For this work data wrangling includes data cleaning, unifying the format for all data streams, and feature extraction. The choice of pre-processing methods is of course application and data dependent. A description of the techniques used for the case study in more details can be found in section 4.

2) DATA STEAM RELATIONSHIP ANALYSIS

a: QUANTITATIVE ANALYSIS

Once data preprocessing is complete, relationship analysis is performed on each pair of data streams to identify those with shared temporal trends. The reduced set of 'related' data streams can then be used to narrow in on potential resilience models. This approach can greatly reduce search time, and computation cost as the initial resilience model hypothesis space for each data stream includes all models covering the range of possible dependencies on all other data streams. For this work, the Granger causality test is used to identify whether one data stream can be used to improve prediction estimate accuracy of another data stream due to shared temporal trends. More specifically, the null hypothesis of no causal relationship is investigated with an F-test, and the resulting p-value is compared to a threshold to identify whether the null hypothesis can be rejected. Here, 'causality' is a misnomer, as the method does not identify causality between data stream sources, and instead implies predictive causality. The method does not take into account the possibility that both data streams are consequences of a common cause, i.e., the existence of latent variables that Granger-cause both data streams of interest.

The Granger test is used rather than a more common correlation metric such as Pearson's product moment as it indicates the statistical significance of using past values of data stream $\mathbf{y}^{\mathrm{m}}$ to assist in predicting $\mathbf{y}^{\mathrm{l}}$ rather than using past values of $\mathbf{y}^{\mathrm{l}}$ alone. Identifying the latter can assist in discovering potential relationships between the data sources, which can be used to help improve resilience models. For this work, bidirectional causal relationships were tested between each pair of data streams. As a pre-processing step, each data stream was first confirmed to be stationary by use of the Augmented Dickey-Fuller (ADF) and Kwiatkowski Phillips Schmidt Shin (KPSS) unit root tests [21]. In evaluating the Granger casual relationships, the lag parameter was programmatically selected using the Akaike information criterion (AIC) [1], ensuring dynamic response of the system.

Identifying the predictive causal relationship between one dependent and two independent variables can be performed using the multivariate Granger causality test which is reliant on the results of VAR analysis. Thus, in the first iteration, the prediction accuracy of all relevant VAR models $\hat{\mathrm{y}}_0^{\mathrm{l}} = \mathrm{f}\left(\mathbf{y}^{\mathrm{l}}, \mathbf{y}^{\mathrm{m}}, \mathbf{y}^{\mathrm{n}}\right)$ can be computed and the field of potential models whittled down for future iterations by subsequent

multivariate Granger analysis. Using this method can greatly reduce the hypothesis search space for resilience models.

b: QUALITATIVE ANALYSIS

Qualitative analysis can be used to determine if underlying latent parameters dictate the relationship between data streams. If such parameters are found, they can be used to reduce the hypothesis space of possible resilient models, thus reducing computation cost and required data sharing network traffic. For this work, the multi-dimensional data scaling (MDS) method was used to visualize the relationship between potential descriptive variables and resilience model performance. MDS operates by mapping points from a high dimensional Euclidian space to a lower dimensional space while attempting to preserve dissimilarity relationships between the points. For the case study, geospatial, topological and demographic parameters are investigated for their utility in predicting resilience model accuracy.

3) MODEL SELECTION AND EVALUATION

The next step is identifying and ranking resilience models by prediction accuracy. For this work, the hypothesis space of resilience models is limited to linear AR and VAR models with one to three independent data stream variables, although this method can be generalized to a larger number of independent variables. Linear AR and VAR models were chosen due to their ease of computation and interpretation for dynamic multivariate time series, as well as their availability on scalable big data platforms. For N data streams, the set of possible models include:

1) N models of type AR using past data from the stream of interest
2) $\left(N^2 - N\right)$ models of type VAR using past data from the stream of interest and a supplemental data stream ('two-city')
3) $(N^3 - 3N^2 + 2N)/2$ models of type VAR models using past data from the stream of interest and two supplemental data streams ('three-city')

The three model types can be expressed by the time series p-th order VAR equation which uses p past data stream values:

$$\hat{y}_t^l = c + \sum_{k=\{l,S\}} \sum_{n=1}^{p} \beta_{t-n}^k y_{t-n}^k, \tag{1}$$

where $\hat{y}_t^l$ is the approximation for the missing data value y_t^l, c is a constant, and β_{t-n}^k is the auto-regression weight learned for data value y_{t-n}^k for data stream k and time t-n. k is summed over the set of data streams to be used in the approximation analysis which includes the data stream of interest l and the set of supplemental data streams S. For model type one, simple AR, no supplemental data streams are used and S is the empty set. Thus, regression is performed over only the past values of the single data stream y^l. For model types two and three, S is composed of the one or two supplemental data streams, respectively. The order of the VAR model used, p, also known as the lag, is dynamically determined by selecting the value of p that provides the minimum AIC.

Model evaluation is performed using time series cross-validation [11], and performance is measured using mean square error (MSE). Time series cross-validation is selected over general cross-validation as it provides better estimates of model prediction performance. For each run of the cross-validation, testing is performed on the value y_t^l, for each possible t and training is made of all possible sets with target y_{t-r}^k with $r \in \{1, \ldots, v - p\}$ and independent inputs y_{t-r-q}^k, $q \in \{1, \ldots, p\}$ with p the lag. MSE is computed over the set of $\hat{y}_t^l$ estimated. Ranking model performance is achieved by comparing the MSE for each model to the AR model MSE for the same target data stream. This emphasizes the improvement in prediction performance provided by the model of interest relative to the baseline of AR. The formula used is:

$$RelMSE = 100 * \frac{MSE\left(f^{AR}\right) - MSE\left(f^i\right)}{MSE\left(f^{AR}\right)}. \tag{2}$$

4) NETWORK DYNAMICS

As trends change in the data streams, the network should respond dynamically, self-adapting and reform network connections to maintain optimal performance. For example, a weather sensor network should respond appropriately as a storm travels from one neighborhood to another. If a sensor experiences data loss, the supporting sensor data used to reinforce resilience should be from those currently experiencing similar weather patterns. Network dynamics are introduced by iterating network evaluation on a user-defined interval, ensuring that network connections reflect current data stream trends. Network re-evaluation is diagramed in Figure 2, starting with data pre-processing, followed by relationship analysis performed for data stream sets, ranking of models by prediction performance and implementing the optimal set of models in the current resilience network. Additionally, if an anomaly in the network is detected, such as the loss of a networked data stream, the network can dynamically select the next best models from the ranks of models to replace those affected. Or in case of an addition of a data stream, a reevaluation of the network is being triggered, and a new list of best models is created. In implementing such a system, a delay may be necessary to improve system stability, reducing the likelihood of rapidly alternating between models due to small variations in data. Network reevaluation can also be triggered based on an external signal ensuring user control or interaction with a relevant event detection system. This depends on domain application of interest, data rates, and sensitivity.

IV. CASE STUDY: CROSS COUNTY CRIME

A. DATA COLLECTION AND PREPROCESSING

For this work, a diverse set of data was collected including crime statistics, weather data, demographics data and geospatial data. Weather data is included as previous research has found that temperature influences crime rates [18].

Scalar geospatial and demographic data are included to explore the relationship between these city parameters and the resilience network. The data characteristics are summarized as follows:

a) *Crime dataset*[2]: 116375 records were collected for crime events reported throughout Montgomery County, Maryland (MD) for the 1/1/2014 to 5/26/2016 period. Each record has twenty-four attributes including date and time (start, end, police dispatch) for the incident, location of the incident (longitude, latitude, zip code, city, state, address, description of urban or residential environment city), police district name and number, agency, uniform crime reporting number and description. For this work, only the incident start date, city, and description fields were used.

b) Weather dataset[3]: Daily weather data was collected over the same period as the crime dataset, for the cities in Montgomery County, MD. Each record has the attributes: temperature, humidity, sea level pressure, visibility miles, wind speed and direction, dew point, precipitation, cloud coverage. Each attribute is described with min, mean, max features. For this work, only the daily mean temperature is used. Also, Montgomery County is covered by three weather centers (College Park Airport, (MD), Ronald Reagan Washington National Airport, (VA) and Montgomery County Airpark, (MD)). For this analysis, data from the Montgomery County Airpark was selected.

c) Census dataset[4]: Census data was collected for the Montgomery County cities, including population count, a number of city dwellers with an education degree of bachelor or higher, and the median household income.

d) Geospatial topological distance: The distances between cities were collected by using Google Maps[5] service, where the distance is calculated as a driving distance from the center of the town to the center of the city.

For this case study, it is assumed that the reporting city collects the event data streams, and each Montgomery County city is identified as a separate data stream and a single node in the resilience network graph. The crime dataset was programmatically preprocessed by first removing crime incidents that occurred outside Montgomery County or crime incidents with unresolvable errors in the location field. Next, the crime incident context features were extracted, and the incidents were sorted by offense type using the MD offense policy categorization, which includes among sixty different categories including assault, abuse, burglary, offense, robbery, theft. Records pertaining to non-crime events, such as natural death, were removed. For each city and date, the number of crime events was then tabulated. A plot of the number of crime events occurring in each city during the last month of data for the available dates of May 1st and May 26th (showing the last 26 out of 877 days) is shown in Figure 3. From this figure, it is evident that Silver Spring dominates of crime events while Chevy Chase, Potomac, and Montgomery Village typically fall near the bottom. with an average of only a few events per day. This trend is shared throughout the period investigated. The eight cities with the highest number of crime events were chosen for analysis, they are (in descending order) Silver Spring, Bethesda, Gaithersburg, Rockville, Germantown, Montgomery Village, Potomac, and Chevy Chase. The rest of the cities were not considered for analysis as they exhibited a deficient number of events per day, e.g., between zero and two daily events. The eight city-based daily crime rate data streams and the one daily weather data stream were normalized by subtracting the mean of each data stream and dividing each data stream by its the standard deviation.

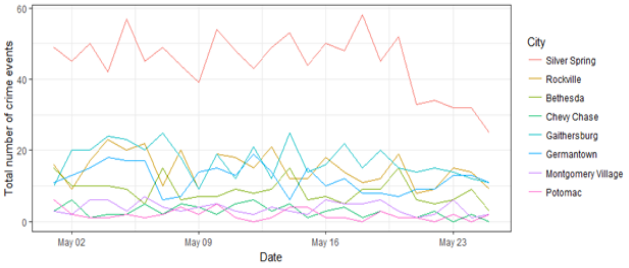

FIGURE 3. Interpreting the number of daily crimes for May 1-26, 2016.

All work described in this paper was performed in the R statistical computing language.[6] The vars[7] package was used for implementing AIC, Granger test, AR models, and VAR models. Graphics were generated using the ggplot[8] package.

B. RELATIONSHIP ANALYSIS: IDENTIFYING POTENTIAL NETWORK CONNECTIONS

1) BIVARIATE AND TRIVARIATE GRANGER CAUSALITY TEST

After data normalization, each crime rate data stream, as well as the weather data stream, were programmatically confirmed to be stationary. The Granger test was then applied to each pair of data streams to quantify the bi-directionally predictive causal relationships, as described above, with the lag parameter automatically selected by the AIC method. Similarly, the multivariate Granger test was performed for each triple of data sources. For both types of models, the weather was removed from the set of target variables. The results of the Granger test analysis for two-city models are shown in Table 1.

A Granger test p-value below or equal to the significance level of 0.05 is used to identify if the forecast data stream is

[2]https://data.montgomerycountymd.gov/
[3]www.wunderground.com
[4]http://www.census.gov/
[5]The identification of any commercial product or trade name does not imply endorsement or recommendation by the NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.
[6]https://www.r-project.org/
[7]https://cran.r-project.org/web/packages/vars/index.html
[8]https://cran.r-project.org/web/packages/ggplot2/index.html

TABLE 1. Granger causality relation index between top eight cities by the number of crime events.

		Forecaster								
	Data Stream	Silver Spring	Rockville	Bethesda	Chevy Chase	Gaithersburg	Germantown	Montgomery Village	Potomac	Weather
Target	Silver Spring	----------	**0.00009**	**0.01685**	**0.03399**	0.17606	0.10106	**0.00925**	0.06527	**0.00796**
	Rockville	**0.02921**	----------	0.09925	0.28008	**0.00313**	**0.03449**	**0.00111**	0.05617	**0.01097**
	Bethesda	0.11235	**0.00658**	----------	**0.01854**	**0.00786**	**0.01437**	**0.00026**	0.21347	**0.02932**
	Chevy Chase	**0.01069**	0.67876	**0.01969**	----------	0.55178	0.85620	**0.04985**	0.05769	**0.03993**
	Gaithersburg	**0.01166**	**0.01822**	**0.00011**	**0.00072**	----------	0.09518	0.06589	0.10123	**0.00426**
	Germantown	**0.01208**	0.35502	0.39609	**0.04996**	0.31938	----------	0.05555	0.77694	**0.01047**
	Montgomery Village	**0.00944**	**0.00173**	**0.04038**	**0.04164**	**0.00491**	0.05677	----------	**0.02273**	**0.00316**
	Potomac	0.07615	0.170987	0.25884	0.24689	**0.01017**	0.06066	0.23479	----------	0.24719

a good predictor for the target data stream. The Granger test indicates that in 57 % of two-city models and 37 % of three-city models the forecast data stream provides statistically meaningful information about future values of the target data stream, and can, therefore, be used to improve prediction of the target data stream. Using the Granger test narrows the hypothesis space from 289 potential models (90 two-city models and 199 three-city models) to 120 models or 42 % of the original hypothesis space. Each indicated Granger-causal data stream pair can now be investigated for prediction accuracy using VAR. It was also found that for all eight cities, weather Granger-causes the daily crime rates either individually or with an additional supplemental data stream, confirming the results from [18].

The resilience performance for all models was computed and compared to the Granger test predictions to identify the efficacy of the Granger test. It was found that the Granger test accurately identifies a predictive causal relationship among 61 % of the two-city models and 61 % of the three-city models. Here, the Granger test is said to accurately identify a predictive causal relationship when the Granger test p-value was equal to or less than the threshold value of 0.05 and the multi-city VAR model provides a lower prediction error than the AR model for the city being predicted. Among the top three models for each city, only two models were misclassified, and these were both the third best models for their cities. Thus the Granger test has an excellent ability to greatly reduce the hypothesis search space while still retaining the best performing models for each city. It was also confirmed that the significance level of 0.05 is optimal in detecting Granger-causation over the range of 0.03 to 0.08 with maximum performance at 0.05.

2) QUALITATIVE RELATIONSHIP IDENTIFICATION

Once a set of resilience models have been selected and analyzed for their performance (described in the next section) in the first iteration, the hypothesis space of possible two-city resilience models can be whittled down for future iterations through qualitative analysis. Qualitative analysis can identify city parameters that may underlie the predictive performance

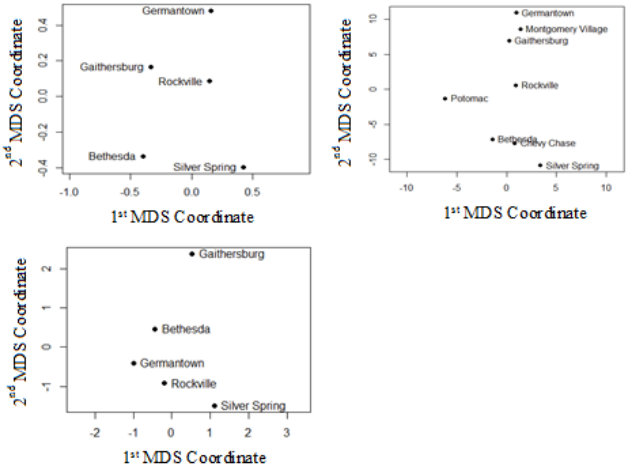

FIGURE 4. Graph representation of cities in Montgomery County, Maryland, U.S.A. by three dimensions: (a) mean square error from model two, (b) distance in miles between the cities and (c) demographics (population, education, and income).

of the resilience models. For instance, if it is found that cities separated by vast distances tend to be poor predictors for each other, a threshold on city-to-city distance can be used to reduce the model hypothesis space. The city parameters investigated include city-to-city distance as well as a set of city demographics including population, the number of denizens with a bachelor degree or higher, and average household income. For this work, MDS is used for qualitative analysis of potential predictive city parameters. First, a two-dimensional mapping is identified for city-to-city dissimilarity, where dissimilarity is defined by the maximum two-city VAR prediction error for each pair of cities (Figure 4. (a)). Here Chevy Chase, Montgomery Village, and Potomac were removed due to their significantly lower crime rates, which result in difficulty comparing prediction error with the rest of the cities. This does not affect the MDS plot as the three cities fall near the origin, and the five other cities retain their relative position. This mapping is compared to the geospatial map of city-to-city distances (Figure 4. (b)). Additionally, each city is described by a vector of the city-based demographics data.

TABLE 2. (a) **Validation metrics, mean squared error (MSE) for model one and two.** (b) **Percentage improvement of model two using model one as a base.**

(a)

		Forecaster								
Target	MSE	Silver Spring	Rockville	Bethesda	Chevy Chase	Gaithersburg	Germantown	Montgomery Village	Potomac	Weather
	Silver Spring	1.11729	1.05356	1.13045	1.09662	1.09721	1.10938	1.09081	1.12080	1.08126
	Rockville	0.95685	0.97789	0.98851	0.96154	0.95159	0.94610	0.99172	0.98421	0.96799
	Bethesda	0.97911	0.97003	0.99561	0.99549	0.96035	1.01674	0.93352	1.00682	1.00749
	Chevy Chase	0.75906	0.75919	0.77145	0.76441	0.77892	0.76088	0.75464	0.75101	0.76816
	Gaithersburg	1.00187	0.94575	0.97302	0.98055	0.97684	0.96209	0.98702	0.96385	0.96160
	Germantown	1.01198	0.99755	1.01942	0.98335	1.00427	1.01820	1.02760	1.03902	1.001856
	Montgomery Village	0.92669	0.90070	0.93184	0.93170	0.89326	0.94129	0.92424	0.93403	0.92309
	Potomac	0.68546	0.67539	0.67157	0.65870	0.64944	0.71280	0.67835	0.66590	0.670914

(b)

		Forecaster								
Target	MSE	Silver Spring	Rockville	Bethesda	Chevy Chase	Gaithersburg	Germantown	Montgomery Village	Potomac	Weather
	Silver Spring	----------	5.70353	-1.17856	1.84972	1.79683	0.70770	2.36989	-0.31464	2.43592
	Rockville	2.15196	----------	-1.08537	1.67231	2.68973	3.25112	-1.41444	-0.64589	0.73619
	Bethesda	1.65704	2.56950	----------	0.01200	3.54149	-2.12269	6.23599	-1.12574	-1.00834
	Chevy Chase	0.69909	0.68260	-0.92130	----------	-1.89847	0.46142	1.27733	1.75175	-0.76476
	Gaithersburg	-2.56224	3.18280	0.39153	-0.37934	----------	1.51011	-1.04254	1.32985	1.42131
	Germantown	0.61066	2.02838	-0.11969	3.42264	1.36901	----------	-0.92275	-2.04451	1.60526
	Montgomery Village	-0.26557	2.54621	-0.82285	-0.80734	3.35167	-1.84473	----------	-1.05966	-0.22616
	Potomac	-2.93622	-1.42485	-0.85108	1.08158	2.47264	-7.04199	-1.86857	----------	-0.75244

TABLE 3. **The best three results from all three models for each data stream and the percentage of improvement compared with model one as a baseline.**

City	M_1	M_2	M_3
Silver Spring	Silver Spring + Rockville + Chevy Chase (1.038511; 7.050432 %)	Silver Spring + Rockville + Montgomery Village (1.044873; 6.481486 %)	Silver Spring + Rockville + Weather (1.046729; 6.315370 %)
Rockville	Rockville + Silver Spring + Germantown (0.9253039; 5.377611 %)	Rockville + Germantown (0.94610; 3.25112 %)	Rockville + Silver Spring + Weather (0.9482259; 3.03348 %)
Bethesda	Bethesda + Rockville + Montgomery Village (0.9177663; 7.818525 %)	Bethesda + Silver Spring + Montgomery Village (0.9289618; 6.69420757 %)	Bethesda + Montgomery Village (0. 93352; 6.23599 %)
Chevy Chase	Chevy Chase + Silver Spring + Germantown (0.7378122; 3.478879 %)	Chevy Chase + Bethesda + Silver Spring (0.7388459; 3.34429167 %)	Chevy Chase + Potomac + Silver Spr ing (0.7402985; 3.15426276 %)
Gaithersburg	Gaithersburg + Montgomery Village + Rockville (0.9369844; 4.080038 %)	Gaithersburg + Rockville + Bethesda (0.9 417804; 3.589083166 %)	Gaithersburg + Rockville (0.94575; 3.18280 %)
Germantown	Germantown + Chevy Chase (0.98335; 3.42264 %)	Germantown + Rockville + Weather (0.9930925; 2.465970 %)	Germantown + Rockville (0.99755; 2.02838 %)
Montgomery Village	Montgomery Village + Rockville + Gaithersburg (0.8883964; 3.877734 %)	Montgomery Village + Gaithersburg (0.89326; 3.35167 %)	Montgomery Village + Rockville (0.90070; 2.54621 %)
Potomac	Potomac + Gaithersburg (0.64944; 2.47264 %)	Potomac + Weather + Gaithersburg (0.6579727; 1.191014 %)	Potomac + Chevy Chase (0.65870; 1.08158 %)

The demographics data is normalized by subtracting the mean and dividing by the standard deviation of each demographic parameter. An MDS two-dimensional mapping is performed using the Euclidean metric (Figure 4. (c)).

It can be seen that the mapping of prediction performance is highly similar to the geospatial mapping, with the cities occurring in similar relative locations except for Gaithersburg. All city pairs also occur at the same relative cardinalities, e.g., in both mappings, Bethesda appears to the left and above Silver Spring. The similarity between mappings indicates that geospatial positioning may be a good predictor for resilience model performance and may also be a good choice of city parameter to reduce the hypothesis space of possible resilience models, with cities that are geospatially

far apart less likely to have high performing resilience models. By restricting the hypothesis space for two-city VAR to only the two nearest neighbor cities among the five towns of interest, the best or the second-best models for each city is captured with the average performance compared to AR going from a maximum of 4.4 % to 3.4 %. Thus, a search space of $N^2 - N$ models can potentially be reduced to $2N$ models.

Investigation of the demographics mapping shows a lower agreement with the resilience performance mapping, indicating that two cities may be more likely to share crime rate trends if they are neighbors than if they share demographic trends.[9] However, these demographics results may be due to the chosen demographics property and used normalization, suggesting further investigation.

C. RESILIENCE MODEL EVALUATION

All possible resilience models were investigated for their prediction performance. The hypothesis space includes all possibilities of the three model types: 1) AR models, 2) two-city VAR models, and 3) three-city VAR models. In our case, we have 9 data streams, and after the data selection functionality, we have the hypothesis space of 9 AR models, 72 VAR models with one supplemental data stream, and 252 VAR models with two supplemental data streams.

Table 2a shows the MSE prediction errors computed for the first two model types, over the full-time range, with forecast data streams listed as columns and target data streams listed as rows. AR models fall along the table diagonal with the rest describing two-city VAR models. Table 2b provides the percent improvement in prediction for the two-city VAR models over the AR models for each data stream. The best model is indicated with the color coding.

For the third model type, three-city VAR, all the approximately one thousand models were evaluated. For simplicity, the top three performing models for each city is listed in Table 3 along with the models' MSE and their percent improvement over the AR model. It was found that three-city VAR models are among the top performing or second best-performing models for each city. As can be seen, for all city data streams the use of additional data sources provides improved prediction and thus improved resilience in the case of data loss. For each city data stream, at least one other source can be used to improve prediction accuracy over simple AR with a maximum improvement of 7.8 %, an average improvement of 4.7 % for all cities, and an average improvement of 5.6 % when excluding the cities with few crime events per day.

The top model for each city is chosen for implementation in the resilience network, see Figure 5. In dynamic operation (discussed in the next section), if an event results in the inability to use the top model, that model is then replaced by the next best model, and so forth.

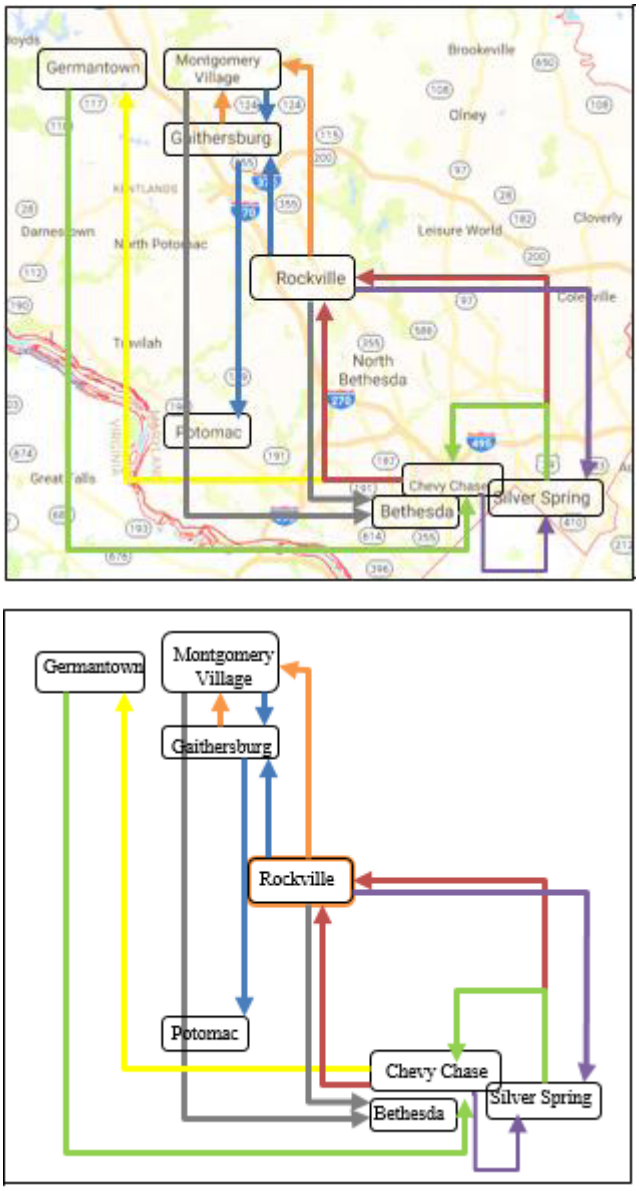

FIGURE 5. Optimal network resilience graph representing data sharing directionality between the cities.

D. RESILIENCE NETWORK DYNAMICS

Resilience network dynamics allow the model to self-adapt to changes in the data streams, so that it always provides optimal performance. Iterating network determination achieves dynamics at user-determined intervals or from a user-provided trigger signal. Figure 6 shows a dynamic implementation for Silver Spring with only models of type one and two investigated. For this implementation, at each date, the network is provided data from the previous four weeks, ensuring that trends learned by the models are local in time. The model which provides the best performance is chosen dynamically for network implementation. Here it can

[9] "Everything is related to everything else, but near things are more related than distant things", First Law of Geography. Tobler W (1970) Economic Geography, 46(2): 234-240.

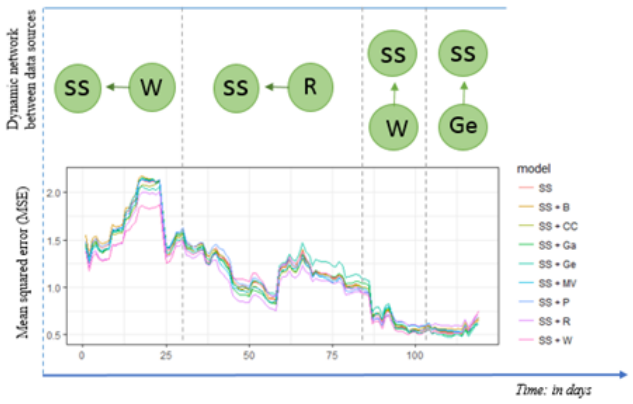

FIGURE 6. Dynamic network implementation for Silver Spring using AR and two-city VAR models. The optimal resilience model initially requires only weather as the independent variable, and switches to Rockville on day 28, weather on day 83, and Germantown on day 106. Legend: Silver Spring (SS), Weather (W), Rockville (R), Bethesda (B), Chevy Chase (CC), Gaithersburg (Ga), Germantown (Ge), Montgomery Village (MV), Potomac (P).

be seen that for the first four weeks, use of weather data provides the best prediction performance. The network graph for Silver Spring is diagramed above these dates, with a directed edge from weather to Silver Spring. On day 28, the optimal resilience model changes, with the weather being replaced with the Rockville data stream. On day 83, the network updates again to depend on the weather data stream and day 106 to the Germantown data stream. As discussed above, in implementing this system delay between model analysis and model selection may be necessary to improve system stability. Selecting the best model with a user-specified periodicity will reduce the likelihood of rapidly alternating between models due to small variations in data. For example, the Montgomery County city-to-city network may be re-evaluated on a weekly basis.

V. DISCUSSION AND CONCLUSION

This paper presents a dynamic network model for improving smart city resilience to data loss. The system utilizes the Granger causality test to identify statistically significant shared temporal trends across multivariate data streams and utilizes VAR to capitalize on those trends to ensure improved data prediction in the case of data loss. Each data stream is provided a ranking of potential resilience models with the top performing model selected for implementation in the network. If the top model can no longer be executed for a particular data stream, the next best model is selected. Iterative evaluation of the system provides a dynamic, self-adaptability to changes in data quality, loss of data streams, and the addition of new information flows.

The network model is demonstrated on City-based daily crime rates reported in eight cities across Montgomery County, MD as well as a daily weather data stream. The optimal resilience network is identified and successfully demonstrated. It is shown that utilizing shared temporal trends between cities provides improved crime rate prediction and resilience to data loss, compared to the use of city-based AR, with a maximum improvement of 7.8 % found in Silver Spring, an average improvement of 4.7 % for all cities, and an average improvement of 5.6 % for cities with high crime rates. The Granger causality test is demonstrated to accurately indicate predictive causality among 61 % of models with all the best performing and second-best models correctly identified as predictively causal. Additionally, the weather is shown to be a top choice for a supporting data stream by both the Granger causality test and VAR performance. This reinforces the finding that weather is a good predictor of crime rates. It was also qualitatively found that small city-to-city distances are a good indicator that temporal trends between city pairs will provide utility in VAR models.

The proposed solution is versatile and applicable to a wide variety of data types and application areas. While demonstrated with an inter-city network, the system can be implemented on other data stream networks such as distributed local clouds in Smart City environment and can be used as an input to recommendation engines sensitive to dynamically changing environments. Due to the use of common statistical methods, this network system can also be scaled on common platforms. Future work on the system will investigate the use of time-varying coefficients in VAR to enhance dynamic performance. The current study will also be expanded to explore the impact of mixed frequency data on system performance. Of additional interest is the use of data stream assigned 'trust' weights, which will allow the user to increase the impact of trusted data sources over those with lower reliability.

REFERENCES

[1] H. Akaike, "A new look at the statistical model identification," *IEEE Trans. Autom. Control*, vol. AC-19, no. 6, pp. 716–723, Dec. 1974.
[2] A. Akbar, F. Carrez, K. Moessner, and A. Zoha, "Predicting complex events for pro-active IoT applications," in *Proc. IEEE 2nd World Forum Internet Things (WF-IoT)*, Dec. 2015, pp. 327–332.
[3] S. Aman, M. Frincu, C. Chelmis, M. Noor, Y. Simmhan, and V. K. Prasanna, "Prediction models for dynamic demand response: Requirements, challenges, and insights," in *Proc. IEEE Int. Conf. Smart Grid Commun.*, Nov. 2015, pp. 338–343.
[4] O. Anava, E. Hazan, and A. Zeevi, "Online time series prediction with missing data," in *Proc. 32nd Int. Conf. Mach. Learn. (ICML)*, Jul. 2015, pp. 2191–2199.
[5] H. Chai and W. Zhao, "Byzantine fault tolerant event stream processing for autonomic computing," in *Proc. IEEE 12th Int. Conf. Dependable, Auto. Secure Comput. (DASC)*, Aug. 2014, pp. 109–114.
[6] W. Derguech, E. Bruke, and E. Curry, "An autonomic approach to real-time predictive analytics using open data and Internet of Things," in *Proc. IEEE 11th Int. Conf Auto. Trusted Comput. (UTC-ATC-ScalCom)*, Dec. 2014, pp. 204–211.
[7] I. S. Dogra and Z. Kobti, "Improving prediction accuracy in agent based modeling systems under dynamic environment," in *Proc. IEEE Congr. Evol. Comput.*, Jun. 2013, pp. 2114–2121.
[8] J. H. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*, vol. 1. Berlin, Germany: Springer, 2001.
[9] M. S. Gerber, "Predicting crime using Twitter and kernel density estimation," *Decision Support Syst.*, vol. 61, pp. 115–125, May 2014.
[10] L. Gunderson and D. Brown, "Using a multi-agent model to predict both physical and cyber criminal activity," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, vol. 4. Oct. 2000, pp. 2338–2343.
[11] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. Amsterdam, The Netherlands: Elsevier, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167923614000268

[12] M. Hvistendahl, "Crime forecasters," *Science*, vol. 353, no. 6307, pp. 1484–1487, 2016.
[13] K. Leong and A. Sung, "A review of spatio-temporal pattern analysis approaches on crime analysis," *Int. E-J. Criminal Sci.*, vol. 9, pp. 1–13, Feb. 2015.
[14] R. Liao, X. Wang, L. Li, and Z. Qin, "A novel serial crime prediction model based on Bayesian learning theory," in *Proc. Int. Conf. Mach. Learn.*, vol. 4. Jul. 2010, pp. 1757–1762.
[15] L. Mookiah, W. Eberle, and A. Siraj, "Survey of crime analysis and prediction," in *Proc. FLAIRS Conf.*, Apr. 2015, pp. 440–443.
[16] T. Nguyên and J. A. Desideri, "Resilience for collaborative applications on clouds," in *Proc. Int. Conf. Comput. Sci. Appl.*, Jun. 2012, pp. 418–433.
[17] S. Pravilovic, M. Bilancia, A. Appice, and D. Malerba, "Using multiple time series analysis for geosensor data forecasting," *Inf. Sci.*, vol. 380, pp. 31–52, Feb. 2017.
[18] M. Ranson, "Crime, weather, and climate change," *J. Environm. Econ. Manage.*, vol. 67, no. 3, pp. 274–302, 2014.
[19] G. W. Schwert, "Why does stock market volatility change over time?" *J. Finance*, vol. 44, no. 5, pp. 1115–1153, 1989.
[20] M. Tokumitsu, K. Hasegawa, and Y. Ishida, "Toward resilient sensor networks with spatiotemporal interpolation of missing data: An example of space weather forecasting," *Proc. Comput. Sci.*, vol. 60, pp. 1585–1594, Jan. 2015.
[21] E. Zivot and J. Wang, "Vector autoregressive models for multivariate time series," in *Modeling Financial Time Series With S-PLUS*, 2006, pp. 385–429.

OLIVERA KOTEVSKA (S'17) received the B.S. degree in academic computer science and the M.S. degree in intelligent information systems from Ss. Cyril and Methodius University, Skopje, Macedonia. She is currently pursuing the Ph.D. degree in computer science with the University of Grenoble Alpes, Grenoble, France.

Since 2014, she has been a Guest Researcher with the National Institute of Standards and Technology, Gaithersburg, MD, USA. Her interest includes complex network analysis in real-world graphs using statistical techniques, knowledge discovery and natural language processing for text analysis and design of smart city, and Internet of Things applications.

A. GILAD KUSNE received the B.S., M.S., and Ph.D. degrees from Carnegie Mellon University. He is a Staff Scientist with the National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, and an Adjunct Professor with the University of Maryland. His research in materials science is part of the White House's Materials Genome Initiative at NIST, a project which aims to accelerate the discovery and optimization of advanced materials. Toward this goal, he integrates machine learning with physics theory, simulation, experiment, and databases to provide live data analysis tools for experimentalists and to develop autonomous experiments for materials discovery. His research also includes leading the development of optimization algorithms for genetics-based cellular sensors and exploring methods for optimizing smart city design. For his work, he has been awarded the NIST Science Data Management and Capabilities Accolade.

His interests include the use of machine learning to accelerate discoveries and optimization in materials science, genetics, and smart city design.

DANIEL V. SAMAROV received the B.A. degree in mathematics and economics from Rutgers University, NJ, USA, and the Ph.D. degree in statistics from the University of North Carolina at Chapel Hill.

He did predoctoral professional engagements in underwriter for the Chubb group of insurance companies; the Teaching Assistant, introductory statistics; the Graduate Research Assistant, Bioinformatics Group, Becton Dickinson Technologies (BDT) (under the supervision of P. Haaland); the Consulting Statistician, Bioinformatics Group, BDT.

He was a Mathematical Statistician with the Statistical Engineering Division, National Institute of Standards and Technology, in 2009.

His professional interests include machine learning, bioinformatics, multivariate statistics, nonparametric statistics, sparse coding, computational statistics, big data, spatial statistics, blind-source separation, and (non-negative) matrix factorization.

AHMED LBATH received the Ph.D. degree in computer science from the University of Lyon, and hold an Habilitation Diriger des Recherches degree. He is a Full Professor of computer science with the MRIM/LIG Laboratory, University of Grenoble Alpes, France and is also conducting research in collaboration with ITL Laboratory, National Institute of Standards and Technology, Washington DC metro area, USA, where he carried out research activities as a Visiting Professor. He is currently acting as the Project Manager coordinating scientific collaborations in the domain of cyber physical systems and smart cities.

He has authored or co-authored several patents, papers in books, journals, and conferences and he regularly serves as co-chair and/or member of several committees of International conferences, journals, and research programs. His research interests include cyber physical human systems, smart cities, mobile cloud computing, recommendation systems, web services, GIS, and software design.

ABDELLA BATTOU received the Ph.D. and M.S.E.E. degrees in electrical engineering from the Catholic University of America. From 2000 to 2009, he was the Chief Technology Officer, and the Vice President of Research and Development for Lambda OpticalSystems, where he was responsible for overseeing the company's system architectures, hardware design, and software development teams. Additionally, he served as a Senior Research Scientist with the Naval Research Laboratory's High Speed Networking Group, Center for Computational Sciences, from 1992 to 2000. In 2012, he served as the Executive Director of The Mid-Atlantic Crossroads GigaPop founded by the University of Maryland, The George Washington University, Georgetown University, and Virginia Polytechnic Institute. He is the Division Chief of the Advanced Network Technologies Division with The Information Technology Laboratory, National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA. He also leads the Cloud Computing Program at NIST.

His research interests are around networking - information centric networking, high-performance IP networks, and cloud computing systems.