A Large-scale Spatio-Temporal Data Analytics System for Wildfire Risk Management

Ziyuan Wang, Hoang Tam Vo, Mahsa Salehi, Laura Irina Rusu, Claire Reeves, Anna Phan
IBM Research - Australia

ABSTRACT

Wildfires have been a significant concern for communities and fire response agencies in many countries. Hence, it is critical to be able to predict the fire risk in a timely and accurate manner and at granular level. However, this requires accessing and processing large amounts of spatial and temporal data from a number of sources in near real-time, while ensuring the immediate availability of risk measurement results. In this paper, we describe a large-scale data-driven system for personalized risk mitigation, fire response's resource optimization and dynamic evacuation planning. It leverages large spatial and temporal datasets to provide predictive analytics in near real-time and to deliver tailored insights to government agencies, communities and individuals.

1 INTRODUCTION

Rapid advances in data acquisition technologies are increasing the volume of data that needs to be processed and stored. The massive amount of data coming from satellites, diverse devices in the Internet of Things (IoT), traffic data using GPS, and meteorological observations opens an opportunity to a deeper understanding of the physical world through big data analytics. In the field of risk analysis for wildfire management, there is steady growth with new models, simulators, visualization tools and new datasets being made publicly accessible. By knowing the risk, government agencies, communities and individuals can be better informed and take appropriate measures to mitigate and prepare for wildfire events.

Current approaches for wildfire risk measurement [8] are insufficient to deal with diverse sets of data. They also do not provide individual level risk measurement and recommend mitigation approaches. Specific big spatial data processing requirements present further challenges. Firstly, traditional databases do not support large geospatial datasets of many terabytes in size, and lack the capability of efficiently indexing and joining various data layers. Secondly, disparate data sources from various services contributes to the time consuming pre-processing of geospatial data. Finally, there is a gap between the algorithms used and the ability to deliver tailored actionable insights to end users.

The focus of this paper is an integrated system that can, on one hand, combine diverse sources of relevant data such as weather, real-time sensor networks, satellite imagery, vegetation, properties information, and others, and apply data analytics and simulation tools to extract insights from the integrated data. On the other hand, this system has to connect to emergency command and control centres to provide up-to-date views of the unfolding risks, and to reach communities and individuals at risk of wildfires, to offer timely personalized advice to mitigate the risk.

The work was initially motivated by requirements from the largest emergency management agency in Australia. It was highlighted that an integrated system is required that would be able to forecast and visualize the wildfire risk at property level, and offer timely insights to the agency to help them prepare to fight the fire by pre-deploying available fire-fighting assets where needed most and re-deploying them as the wildfire risk in the area changes.

We describe the capabilities of our system for wildfire risk management with three relevant applications, 1) personalized risk mitigation, 2) dynamic evacuation routing (that would benefit to communities and individuals), and 3) resource pre-deployment optimization (that would benefit to fire emergency agencies).

In our system, the underlying risk model learns from history and previous actions by the application users in order to recognize patterns and to mitigate the risk proactively. The risk analytics can be resolved to the level of an individual property or road segment within a fire danger zone, with insights relevant to the homeowner, evacuees, and emergency agencies. Most importantly, the system is able to:

- Support near real-time and large-scale applications by integrating data from various data sources with different formats, projections, and resolution, and handle information heterogeneity.
- Abstract fundamental building blocks for accurate and time efficient predictive analytics and deliver tailored insights to end users via service REST APIs.
- Reuse and extend the above analytics building blocks to support personalized risk mitigation, resource optimization, and dynamic evacuation planning.

The rest of the paper is organized as follows. Section 2 presents an overview of our system architecture and highlights several analytics components. Section 3 provides three case studies supported by our system in wildfire risk management. Section 4 discusses the future directions and the paper is concluded in Section 5.

2 SYSTEM ARCHITECTURE OVERVIEW

Most of existing big spatial data management and analytics systems [1, 5] are developed on top of Hadoop MapReduce, which has significant overhead when executing iterative algorithms that many machine learning applications rely on. Further, although the system proposed in [5] enables automatic data download, data curation, and scalable storage for big spatial data, it rasterizes all data layers to handle data download, re-projection, and data indexing,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
GeoRich'17 , Chicago, IL, USA
© 2017 Copyright held by the owner/author(s). Publication rights licensed to ACM.
978-1-4503-5047-1/17/05...$15.00
DOI: http://dx.doi.org/10.1145/3080546.3080549

thus limiting the capabilities of vector data processing. Vector data processing is important in our system to support risk prediction and dynamic evacuation routing.

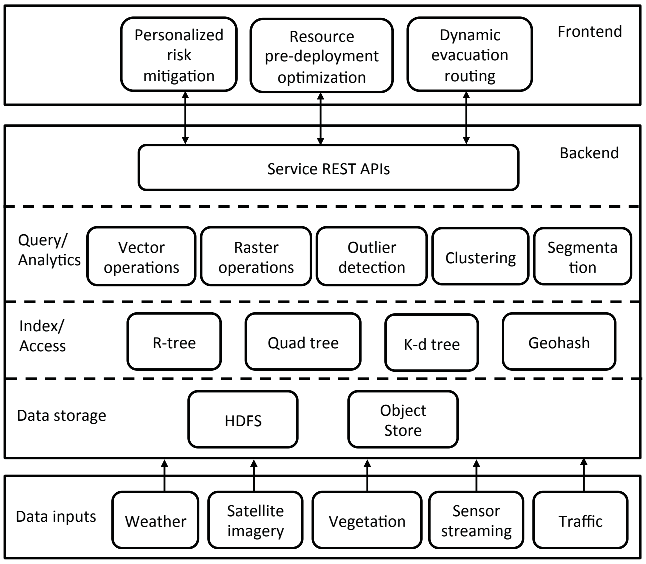


Figure 1: Architecture of a big spatial analytics system for disaster management.

In contrast to Hadoop-based systems, we build our system on top of Spark [12] considering its advantage over Hadoop in running iterative algorithms, due to Hadoop's significant overhead associated with each iteration. We use Spark to access and process both historical and real-time streaming data in a fast batch way. Spark's core data abstraction Resilient Distributed Dataset (RDD) propels a variety of compute-intensive tasks including interactive queries, streaming, machine learning, and graph processing.

Our system consists of the backend and the frontend as shown in Figure 1. The backend is built on Cloud infrastructure [1], which allows the application to be accessed from anywhere. The frontend can be any client, for example, iOS or Android mobile app, and web applications. The backend and frontend communicate via the Service REST APIs. The backend consists of the following layers: data storage, index/access, query/analytics. The three frontend applications are summarized in Figure 2 and presented in Section 3.

We adopt a global-local indexing strategy to achieve better performance. The global index is in the memory of the master node to keep the boundaries of partitions and enables the system to prune irrelevant partitions. The local index is in each partition. In the partition phase, we use the Sort-Tile-Recursive (STR) algorithm [6] similar as in SpatialHadoop [3]. We implement *multiple indexing* to handle diverse datasets because each of the indexing has its own merits. For example, we use R-tree to handle the skewness of road network by adjusting the partition size so that each partition contains similar amount of data. To filter the information that only matters to a user we adopt uniform grid index to spatially co-partition the datasets of vegetation and property address in order to facilitate parallel computing model supported in Spark.

[1] IBM Bluemix. www.ibm.com/cloud-computing/bluemix/

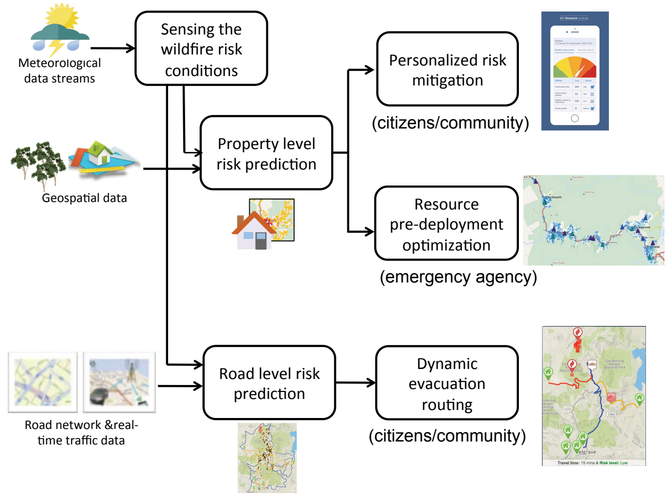


Figure 2: Applications on dynamic risk prediction.

The geospatial capabilities supported in our system are shown in the query/analytics layer in Figure 1. For example, the *vector operations* include: 1) Distance: Given a point and a set of polygons, find the polygons that are within a distance to the point. The distance could be either straight-line or road-network distance. For example, in personalized risk prediction, for each property, we need to find the vegetation polygons nearby. For fire fighting resource optimization, we need to compute the dynamic road-network distance between the moving fire fighting asset and a set of properties. 2) Direction: Given a point and a set of polygons, find the polygons that are within a range of direction to the point. For example, we need to get which set of vegetation polygons would impact a property by the direction of wind. In the dynamic evacuation routing application, a vehicle route needs to avoid the fire spreading direction.

Another component in the query/analytics layer is the *online outlier detection* which is used to identify high wildfire risk conditions. To develop such a model, one of the big challenges is lack of label information for building supervised predictive models. In fact, we noticed that in case of wildfire risk, it is really not reasonable to identify which periods of time are risky and which of them are not. To deal with this challenge an unsupervised algorithm is proposed for wildfire risk prediction [10], where high risk periods are identified without any statistics or expert knowledge. An outlierness value is computed for the incoming streams of weather observations and these outlierness values are considered as wildfire risk of observations. Since the proposed algorithm in this component is completely unsupervised, it can be easily applied in other areas of interest, using other big weather datasets. More specifically, the historical data stream is divided into data chuncks (windows of observations). A near linear clustering-based anomaly detection technique is used to profile the weather observations for each window. Thereafter, the similarity between the clustering profile of current window and the profiles of all historical windows are computed. As a result, more relevant historical profiles are only selected to decide on the outlierness values of current observations. For each incoming observation p, an outlierness value is computed using the

Mahalanobis distance of p and all relevant clusters. Finally, an ensemble of relevant (weighted) historical models is used in assigning final anomalous values to the current observations. One advantage of this analytical component is its low computational complexity which is near linear with respect to the number of weather observations. Hence, the queries can be addressed in near real-time. Also, being an ensemble-based approach is another advantage of this component which makes it highly scalable by parallelization.

3 CASE STUDY

As discussed in the Section 1, we now describe three applications of our big spatial system for wildfire risk management, where the study area is Blue Mountains, New South Wales, Australia. Part of this work was demonstrated previously in [9]. The area was selected because of the availability of data to test our proposed system, and to validate the results with domain experts from an emergency management agency in NSW. As the above mentioned three applications rely on a number of risk prediction steps (see Figure 2), we first describe the raw data inputs as well as the data processing tasks that we applied in order to support the applications.

3.1 Raw Data Inputs

As input for sensing wildfire risk conditions, we choose to conduct a high-resolution meteorological re-analysis (WRF-ARW) [2] that resolves hyper-local phenomena with detailed data, including temperature, precipitation, pressure, relative humidity and wind speed. The study area (Blue Mountains, NSW, Australia) covers an area of $275km \times 275km$ centred on the town of Blackheath (-34.80, 138.90) and has a grid spacing of $5km$, resulting in a 55×55 grid and roughly 3000 grid cell observations. The simulation output is at 1-minute frequency, hence a total of approximately 65,000,000 observations to analyse.

For the Property level risk prediction and Road level risk prediction, the input data is summarized in Table 1.

3.2 Data Processing

The data processing includes loading data into a distributed file system, re-projecting geographic coordinate system where required, transforming data format if necessary, partitioning and building indexes at local and global level, as well as components shown in the query/analytics layer in Figure 1. As an example, Figure 3 illustrates the data processing sequence for dynamic risk prediction, along with data format and projection.

3.3 Personalized Risk Mitigation

The value of quantifying wildfire risk is well established, as it enables government agencies, communities, and individuals to understand and proactively manage the risk.

Property level wildfire risk can be seen as a local process, where attributes such as surrounding tree cover and other vegetation are location-specific and critical for an accurate risk rating measurement. We calculated a series of attributes that are physically based and quantifiable, including shortest distances between property addresses and the adjacent bushland, surrounding tree coverage, and local slope. To obtain these variables, we used the most detailed geospatial datasets available, such as geocoded street addresses,

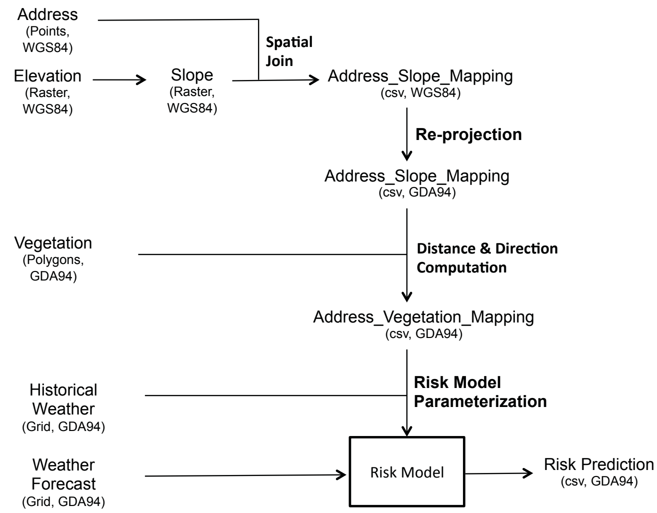


Figure 3: Data processing for dynamic risk prediction.

$5km$ spatial resolution weather data and $25m$ resolution digital terrain models. These variables together with information provided by the user, such as building and roof material, feed into our risk model to calculate a composite risk rating for each property address.

More specifically, the implemented model for wildfire risk is a combination of a series of modules relating to different aspects of wildfire risk, with the relative weights of different aspects having been qualitatively determined. Further development of the modules and quantitative parameterization is presently being undertaken. The modules incorporated in the presented case study are shown in Figure 4, namely the Fire Danger Index [7], Wilson House Survival model [4], a wind vector factor, bushland supporting factor, and Ignition Likelihood Index [11].

The rationale for a modular approach to the fire risk is to support the use of model blending techniques, which enable the customization to local factors, and to support a plug-and-play approach that facilitates continued development of our understanding of wildfire risk.

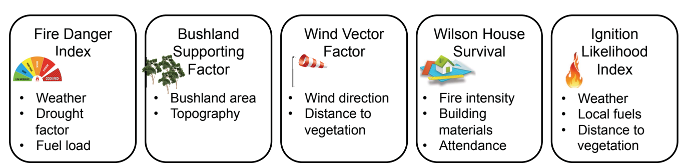


Figure 4: Modules of personalized wildfire risk model.

The results of this model are risk values for each property and for each timestamp in the study period, as exemplified in Figure 5 for households in Blue Mountains, Australia. For risk visualization we use a range of colours, from light blue for lowest risk to dark red for highest risk (see Figure 5 (c) for the full legend).

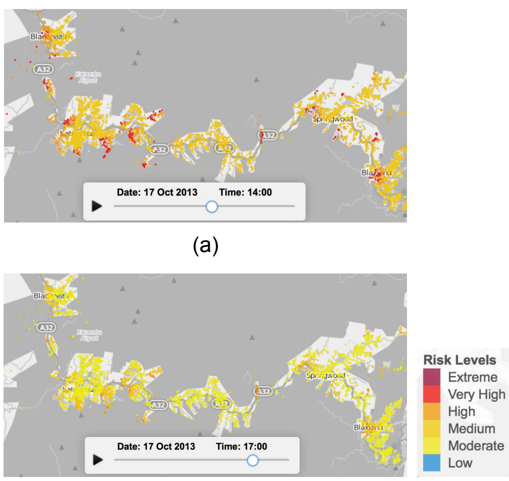
Based on the Property level risk prediction, we build a mobile app that enables individuals to take proactive actions to reduce the risk to their homes. The mobile app provides personalized recommendations to the user on actions they could take to reduce their risk rating, specific to their particular profile. The user is able

Table 1: Overview of input data to the risk model - Blue Mountains

Data	Source	Type/Format	Resolution	Update Rate
Address	Open	Point/Shapefile	NA	Static
Elevation	Open	Raster/GTiff	25 m	Static
Vegetation	Open	Polygon/Shapefile	NA	>Yearly
Weather	WRF-ARW	Grid	5 km	minute
Drought Factor	ADFD	Grid/NetCDF	3-6 km	Seasonally
House specs	User	JSON	NA	Semi-static
Traffic GPS	User	Point/JSON	NA	minute
Road network	OpenStreetMap	Vector/OSM	NA	Semi-static

to see how the wildfire risk might be reduced by taking corrective actions that lead to physical changes to the contributing factors of the risk model. Figure 6 shows two screenshots of the mobile app designed to help people reduce risk via (a) long-term and (b) short-term actions. Once people select actions that could reduce their risk, the system responds in real time with a recalculation of the predicted risk.

From our extensive testing of the system response time, the risk model is able to return the risk assessment at property level in less than a second. Updating the risk assessment of all 43389 properties in the case study area whenever a new weather forecast is received is also performed in less a minute.



(a)

(b) (c)

Figure 5: Visualization of dynamic risks for properties in Blue Mountains at (a) 2pm and (b) 5pm.

3.4 Resource Pre-deployment Optimization

Using the forecast risk values from the previous module, our system enables emergency agencies to take proactive actions to prepare resources to defend communities.

We deploy an optimization tool that assists in making rapid and effective decisions for dynamic reallocation of fire-fighting assets, with the aim of minimising overall fire risk remaining in the targeted area at any given time. Given a risk measure for each individual property for each hour interval, our objective prioritizes coverage of properties with highest risk. Also, response time for resource vehicles is calculated as travel time along the road network.

(a) (b)

Figure 6: Mobile app of personalized recommendation for risk mitigation.

We deal with the complexities of big data computations by developing a *scalable clustering model* that groups similar properties based on the property spatial location, property risk and distance to accessible roads. The clustering model which is based on k-means algorithm, changes every hour to cater for dynamic changes of property risk due to changing weather conditions that lead to relocate the fire-fighting assets during the day. Experimental results show that our scalable optimization model is effective in terms of property coverage and risk reduction while remaining efficient in the amount of time required to obtain solutions, making it practical for real-time applications. The optimization tool uses data from the clustering model and road networks together with information about available fire fighting vehicles and stopping locations. For each hour interval, all available fire fighting vehicles are assigned to stopping locations such that the configuration of vehicles maximizes coverage and the estimated risk reduction across the whole targeted area.

Real world constraints are introduced into the optimization tool to ensure that solutions for assigned vehicle locations are feasible. Travel time between the location of any vehicle at the beginning of an hour, and its assigned destination for that hour interval, must be less than an hour so that it can reach the desired stopping location and provide coverage prior to the system being reoptimized at the next hour interval. Stopping locations include, but are not limited to, designated fire stations. Other suitable locations within the targeted area, for example parks with sufficient parking space and facilities, may also host fire fighting vehicles but are not guaranteed to be open for fire fighting vehicle pre-deployment at every hour of every day. The optimization tool allows the temporal availability of stopping locations to be respected when optimizing pre-deployment decisions. The capacity of each stopping location, that is the number of vehicles that may be situated at the stopping location at the same time, is also pre-determined and constrains the assignment of vehicles in the optimization tool.

Multiple types of fire fighting vehicles are considered: heavy tankers, light tankers and strikers. Different vehicle types have different accessibility to roads throughout the road network and provide a varying degree of risk reduction. For example, heavy tankers carry additional equipment and therefore reduce risk more than either light tankers or strikers but are unable to access certain roads outside of urban areas or must travel at lesser speeds than smaller fire fighting vehicles. Risk reduction is a function of the capability to reduce risk of each fire fighting vehicle within a coverage radius of 10 minutes travel time, as shown in the example in Figure 7. Vehicle placements are more effective at reducing risk when there is a shorter travel time to reach the property at risk. Multiple vehicles can also contribute to risk reduction for a single property, however, the risk of a property cannot be reduced below a lower bound that is greater than zero.

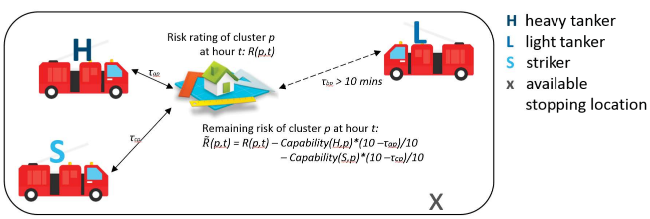


Figure 7: Example of remaining risk after pre-deployment of fire fighting vehicle configuration

The optimization tool utilizes IBM ILOG CPLEX Optimization Studios 12.2 to obtain solutions. The amount of time required to find optimal locations for fire fighting vehicles scales with the number of vehicles available for defensive pre-deployment. For small numbers of fire fighting vehicles, the optimization tool can return results quickly enough to be implemented in response to changes in risk each hour. Testing shows that assigning stopping locations to a configuration of 3 heavy tankers, 5 light tankers and 2 strikers can be solved to 0.01% relative optimality gap with an average solve time of 5 minutes for a full hour interval. For larger fleets of fire fighting vehicles, the optimization tool is useful to pre plan defensive vehicle movements with predicted risk data. Figure 8 visualizes the reduced risk values after pre-deploying 20 compared with 10 fire-fighting assets in the study area. A substantial reduction in the risk of property damage can be noticed, and the range of colours used to visualize risk after pre-deployment is the same as in Figure 5.

3.5 Dynamic Evacuation Routing

During a disaster event it may be difficult for people in the affected areas to make smart decisions about the best routes that are both fast and safe to evacuate, due to psychological and emotional impact of these highly stressful situations. Even worse, affected people may end up moving from one dangerous area to another if they

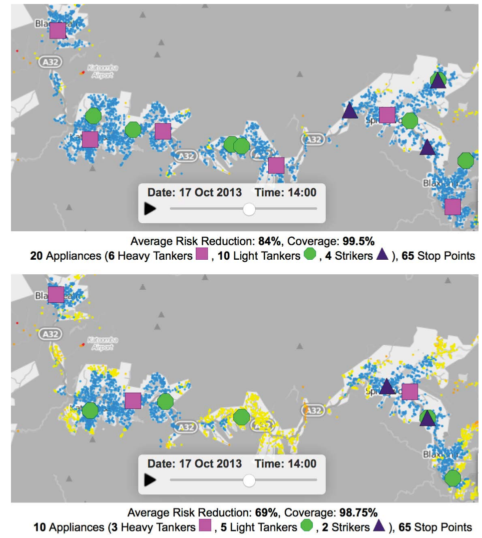


Figure 8: Risk reduction of resource pre-deployment optimization.

are not provided with real-time information about the areas at risk, as situations change during the event. For example, wildfires can unpredictably change direction due to unexpected weather conditions, reaching areas that were initially deemed safe. In this context, the safety as well as timeliness of dynamically updated evacuation routes is extremely critical in responsive management applications and can save lives.

Hence, we develop an evacuation route planner that dynamically suggests the fastest and safest routes that are adapted to the changing situations of emergency events in real-time. The route recommendation algorithm executes on a graph data structure created by the road network. Road segments essentially constitute edges in the transportation graph, while intersections of road segments represent nodes (vertices) of the graph. More importantly, this transportation graph is time-dependent as the costs of the edges change along the time given new situations such as real-time traffic events and up-to-date weather data. Another special characteristic of this transportation graph is that each of its edges is associated with multiple types of costs including travel time and safety when traveling along this edge. We estimate the risk when traveling along a road segment at a given time based on its proximity to the spreading bushfire at that time. Furthermore, as long as most of the vehicles during this evacuation crisis are using GPS-equipped devices, their real-time locations can be collected. Our system estimates the average travel time along a road segment at runtime based on this real-time traffic data.

The recommendation workflow works as follows. First, an evacuee sends a request including their origin location to the system using a mobile application. Then, the system runs an optimization algorithm to calculate the best evacuation route that minimizes evacuation time as well as the risk of being in danger while along the evacuation path. The requests from other users and their suggested routes are also taken in account in the planning for new requests to deal with congestion problem. Finally, the best calculated route is returned to the evacuee and visualized by the mobile application. When the conditions change (e.g., a forecast that predicts a change in wind direction or the traffic data is updated), the system updates the results and alerts the user about the new best path. In addition to real-time recommendation of safest evacuation routes, the system can also provide the affected people with other useful information such as real-time traffic situation as well as potential areas at most risk.

Figure 9 compares the routes of evacuation from a high risk area caused by a wildfire event (red icon) to one of the shelters (green icon). Figure 9 (a) is the fastest route returned by classical shortest/fastest path algorithms. This route is safe at the time being as it is not threatened by the current spreading fire. However, while the evacuee is on the way, a new fire starts in the vicinity of his intended route as shown in Figure 9 (b), now this route is shown in red because it is unsafe anymore. Instead of the static approach, the routing planner is aware of the occurrence of the second fire, as well as the blocked road section due to an accident as shown by the red marker. Therefore, the routing planner returns a new safe route in blue.

(a) (b)

Figure 9: Dynamic routes recommended by the evacuation routing planner.

4 FUTURE WORK

We have identified at least three areas of future work in order to extend the applicability of the proposed system.

User experience: We intend to automate data gathering for sources where currently data inputs are highly manual, for example retrieving house profile information from the user's insurance data rather than user input. IoT devices can provide a variety of personalized data about the house and environment, as well as social media can be used as a rich data source. Another aspect of user experience is to offer automated and near real-time risk assessment with changing weather conditions, without the need for user-driven re-assessment initiation.

System scalability: System scalability is critical in deploying the applications to the users. As communities grow and migrates to different areas, there will be more people living in wildfire-prone areas that do not have any prior knowledge of wildfire risk. Therefore, there is an ongoing need to educate, inform and find new ways of sharing local information with these communities. A mobile app would serve as a community education tool to improve individual wildfire preparedness. To reach the wider population, we envisage making the application available through App/PlayStore, as well as publishing via the relevant government agencies.

Wider applications: While we have applied our system for wildfire scenarios, it can be applied in other areas such as personalized recommendation for flood mitigation in disaster prone areas; individualized risk prediction for beach safety and so on. While extending the applications of our system, we aim to bridge the gap between research and practice. The disaster management industry needs to rely more on science and data than on opinion, so it will be great to demonstrate a suit of risk data and modelling that are applicable in a wide variety of contexts. More evidences from experiments with large datasets across multiple scenarios would also then input into well informed policy decision-making.

5 CONCLUSIONS

Using big spatial and meteorological data with high volumes, velocity and variety to solve real world problems is often complex and challenging. In this paper, we presented an integrated disaster management system that addresses the challenges of handling large-scale datasets and provides predictive risk analytics tailored to individual circumstances. The application of our system can be extended to a wide number of contexts, while allowing personalized use for general public, optimization and policy decisions by disaster management agencies and the government. By continuing to evolve our system with more dynamic data (e.g., sensor data and satellite imagery) and embedding more advanced analytics into the system we will increase accuracy and timeliness of disaster predictions to save lives and assets.

REFERENCES

[1] Xin Chen, Hoang Vo, Ablimit Aji, and Fusheng Wang. 2014. High performance integrated spatial big data analytics. In *BigSpatial*. 11–14.

[2] DP Dee, SM Uppala, AJ Simmons, Paul Berrisford, P Poli, S Kobayashi, U Andrae, MA Balmaseda, G Balsamo, P Bauer, and others. 2011. The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society* 137, 656 (2011), 553–597.

[3] Ahmed Eldawy and Mohamed F Mokbel. 2015. SpatialHadoop: A MapReduce framework for spatial data. In *ICDE*. 1352–1363.

[4] Wilson A. A. G. 1988. A simple device for calculating the probability of a house surviving a bushfire. *Australian Forestry* 51, 2 (1988), 119–123.

[5] Levente J Klein, Fernando J Marianno, Conrad M Albrecht, Marcus Freitag, Siyuan Lu, Nigel Hinds, Xiaoyan Shao, Sergio Bermudez Rodriguez, and Hendrik F Hamann. 2015. PAIRS: A scalable geo-spatial data analytics platform. In *BigData*. 1290–1298.

[6] Scott T Leutenegger, Mario A Lopez, and Jeffrey Edgington. 1997. STR: A simple and efficient algorithm for R-tree packing. In *ICDE*. 497–506.

[7] Alan Grant McArthur. 1967. Fire behaviour in eucalypt forests. *Comm. of Australia For. Timber Bur. Leaflet No. 107* (1967).

[8] Carol Miller and Alan A Ager. 2013. A review of recent advances in risk analysis for wildfire management. *International journal of wildland fire* 22, 1 (2013), 1–14.

[9] Laura Irina Rusu, Hoang Tam Vo, Ziyuan Wang, Mahsa Salehi, and Anna Phan. 2016. Data-Driven Prediction and Visualisation of Dynamic Bushfire Risks. In *Proc. of ADC*. 457–461.

[10] Mahsa Salehi, Laura Irina Rusu, Timothy Lynar, and Anna Phan. 2016. Dynamic and Robust Wildfire Risk Prediction System: An Unsupervised Approach. In *Proc. of SIGKDD*. 245–254.

[11] KG Tolhurst and KA Howlett. 2003. House Ignition Likelihood Index-An Hazard Assessment Method for Land Managers in the Wildland-Urban Interface. In *10th AFAC Conference and 4th International Wildland Fire Conference*.

[12] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J Franklin, Scott Shenker, and Ion Stoica. 2012. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proc. of NSDI*.