IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 June 2022, accepted 2 July 2022, date of publication 12 July 2022, date of current version 18 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3190399

RESEARCH ARTICLE

Correlating Extreme Weather Conditions With Road Traffic Safety: A Unified Latent Space Model

JACOPO FIOR, (Member, IEEE), AND LUCA CAGLIERO, (Member, IEEE)
Politecnico di Torino, 10129 Turin, Italy

Corresponding author: Luca Cagliero (luca.cagliero@polito.it)

This work was supported by the Smart-Data@PoliTO Center for Big Data and Machine Learning Technologies.

ABSTRACT The presence of extreme weather conditions is known to expose drivers to a higher risk to incur in road accidents. Quantifying the correlation between adverse weather conditions and road traffic safety is useful for several reasons such as planning preventive actions, managing vehicle fleets, and configuring alerting systems. However, since the risk of road accidents occurrences within a specific spatial region is influenced by several factors other than the weather conditions, quantifying the actual impact of adverse weather phenomena regardless of the effect of weather-unrelated conditions can be challenging. To tackle the aforesaid issue, this paper proposes to adopt a unified latent space model based on time series embeddings. Firstly, it encodes a subset of historical series reporting weather-related accident occurrences in specific risky areas into the high-dimensional vector representation. It also encodes the weather element measurements acquired by meteorological stations spread over the analyzed area. Then, to estimate the risk level of each region within the same spatial context it seeks the temporal risk patterns that are most similar to those observed in risky areas. The experiments carried out in a real case study confirm the applicability of the proposed approach.

INDEX TERMS Time series embeddings, weather data analysis, road accidents analysis, key performance indicators.

I. INTRODUCTION

Understanding the causes of road accidents is of great importance in several domains, among which urban mobility management [1], preventive maintenance [2], and emergency management [3]. Lots of efforts have been devoted to studying the factors influencing the likelihood of road accidents occurrences. In this regard, traffic and weather characteristics are known to have a strong influence on road traffic safety [4], [5].

The increasing availability of weather and traffic data has fostered a deeper exploration of both traffic characteristics [6]–[8]) and weather conditions (e.g., [9]–[11]). For example, regarding weather effects, the effects of precipitations, temperature, and visibility conditions are known [12]. Weather conditions have been found relevant in the prediction of travel time [13] and fuel consumption [14]. Under the same umbrella, previous studies (e.g., [15]) have highlighted the correlation between the presence of extreme weather conditions (e.g., heavy rain, very low temperatures) and the occurrence of road accidents. The present work specifically studies the correlation with such extreme weather conditions by means of a data-driven approach.

The purpose of this work is to quantify the influence of extreme weather effects on road accidents occurrences within specific spatial regions. Despite previous works have clarified their big influence, two related issues still deserve attention: (1) The likelihood of observing traffic crashes is likely to be influenced by factors other than the occurrences of extreme weather conditions as well [16]–[18]. (2) Statistics about traffic crashes are typically incomplete and characterized by a certain level of uncertainty [19]. For these reasons, in practice, it is difficult to quantify the actual effect of extreme weather conditions on road accidents occurrences regardless of the effect of other conditions.

Motivated by the lack of ad hoc data-driven solutions to the aforesaid problem, we propose to leverage the weather element measurements acquired by meteorological stations

The associate editor coordinating the review of this manuscript and approving it for publication was Dost Muhammad Khan.

within delimited spatial areas where actual weather-related road accidents are reported. Specifically, we present a unified latent space model, based on time series embeddings, that encodes the key information about (1) the weather-related road accidents that occurred in risky areas, and (2) the weather element measurements acquired by meteorological stations in all the analyzed regions (both the risky areas and the unclassified ones).

For our purposes, we denote a context as a set of spatial regions with similar characteristics (e.g., same urbanization level, similar elevation, same geographical district). We estimate the risk level associated with each region within a specific context as the similarity level, computed in the unified latent space, between the weather element measurements acquired by meteorological stations in the region and those observed in risky areas. Since weather measurements are most likely to be acquired with widespread coverage, they can be profitably exploited to identify sequences of adverse weather events that are similar to those that occurred in risky areas, i.e., in the areas where weather-related road accidents have been previously reported.

The presented data-driven methodology consists of three main steps. First, the raw weather measurement series associated with each weather monitoring station are transformed into a set of temporal sequences of observed events describing adverse weather conditions (e.g., very high or very low temperature values, heavy rain). Next, the extracted sequences are embedded into a high-dimensional latent space model to capture similar trends in the series of weather element measurements. Finally, the estimation of the risk levels is performed using four different risk models, which leverage the encoded information using alternative strategies.

To empirically analyze the results achieved by the proposed approach we run a set of experiments on real data acquired from the U.S. Historical Climatology Network and the U.S. Census. The outcomes of the simulations confirm the applicability of the proposed approach in real scenarios where the estimated risk levels can be profitably exploited as Key Performance Indicators to support decision-making.

The paper outline is reported below. Section II overviews the related literature. Section III presents the proposed methodology. Section IV presents the data under consideration and summarizes the main empirical findings. Section V draws conclusions and discusses the future research directions.

II. RELATED WORKS

Prior works have already investigated the effect of weather on road traffic safety. A preliminary attempt to study the effects of weather conditions on daily crash counts was made in [16]. The authors applied auto-regressive time series forecasting models on historical crash data, meteorological data, and traffic exposure data. The idea behind it was to predict the number of car crashes that are likely to occur in the upcoming days by discovering predictive trends from past series of correlated measurements. Similar to [16], the work presented in [15] focused on studying the correlation between the time series of quantitative weather measurements and the counts of injury accidents reported on a monthly basis. In parallel, they also investigated the use of daily time series to study the within-the-month variability of extreme weather effects. A more robust machine learning-based approach to accident count prediction is given in [18]. Recent works also addressed the study of the separate impact of climate and non-climate variables on fatal traffic accidents [20], [21]. The study of accident contributing factors has been enabled by several digital technologies including AI, IoT, and vehicle networks [22], [23]. A systematic review of the most recent applications can be found in [24].

The present work studies the effects of extreme weather conditions on road traffic safety using an unsupervised approach based on time series embedding techniques. Unlike [15], [16], [20], [21], it focuses on estimating the risk level of a spatial context by identifying and comparing risky patterns that occur in the sequences of extreme weather events.

Partly related studies tailored to specific application scenarios have been presented in [25]–[27]. Specifically, in [25] the authors aimed at forecasting crashes on freeways due to reduced visibility. The predictive models achieved around 70% precision in crash identification. The work presented in [26] analyzed the impact of rainfall on road traffic accidents in urban areas, whereas in [27] the authors proposed a crash prediction model to forecast hourly crash likelihood of highway segments. The aforesaid studies highlighted the importance of real-time contextual information such as weather, road surface, and traffic conditions. Furthermore, in [27] the authors also argued that rainfall quantification acquired by meteorological stations is likely to be not representative enough of the actual road safety. The methodology presented in this paper is neither tailored to a specific mobility context (e.g., highways, urban areas) nor to a particular adverse weather condition (e.g., rainfall, low visibility).

Recently, the authors in [17], [18], [28] explored the use of Machine Learning models to forecast road accidents occurrences. Specifically, they trained an ensemble of tree-based models on multivariate, fine-grained weather datasets. Thanks to the inherent interpretability of the predictive models, they have shown that a subset of the considered weather variables is highly discriminating for predicting accident occurrences. This paper focuses on quantifying the correlation with adverse weather conditions to estimate per-context risk levels. The work presented by [19] explored the data sources that have been previously used in literature to study road traffic safety. The main takeaway from the above-mentioned research study is that the presence of data uncertainty may hinder the application of the previously proposed data-driven methodologies. Inspired by the latest research findings, we aim at overcoming the limitations of existing approaches due to the lack (or uncertainty) of weather-related accident data.

A preliminary version of the present work was presented in [29]. It provides a high-level overview of the problem and a qualitative evaluation of some preliminary results. This work substantially extends [29] as follows: (1) It formalizes the problem under analysis (see Sections III-A, III-B, and III-C), (2) It details the procedure used to compute the per-context risk levels (see Algorithm 1), (3) It presents four new risk models aimed at effectively computing the risk level of a region within a given context (see Section III-D), (4) It presents a quantitative strategy to compare the performance of different risk models (see Section IV-D).

III. PRESENTED METHOD

The proposed methodology, namely *Weather Influence on Road Accidents* (WIRA, in short), focuses on quantifying the effect of adverse weather conditions on road accidents occurrences in different spatial contexts.

A pseudo-code of the adopted procedure is reported in Algorithm 1. The main WIRA architectural blocks are enumerated below.

- **Context definition:** it maps geographical areas characterized by specific combinations of geographical and census feature values to different contexts. Each context is then annotated with the locations of the previously reported weather-related road accidents (see Section III-B).
- **Weather data acquisition and preparation:** the raw weather element measurements are acquired by the meteorological stations, collected in a centralized repository, and processed to extract the temporal sequences of adverse weather events. To combine weather data with road accidents counts, each meteorological station is also mapped to the nearest region within the analyzed area (see Section III-C).
- **Time series embedding:** a high-dimensional vector representation of the per-station series of adverse weather events is inferred using an ad hoc neural network-based embedding approach from weather data (see Section III-D).
- **Contextualized risk model:** it quantifies the risk level separately for each of the previously defined contexts. Risk level estimates are based on domain-specific data-driven models designed on top of the unified latent space. (see Section III-E).

A more thorough analysis of each step is reported in the following sections.

Algorithm 1: The WIRA Methodology

```
Input : $\mathbf{A}$: geographical area under consideration;
        $\mathbf{C}$: set of relevant contexts;
        $\mathbf{D_w}$: dataset including weather data relative to all the stations;
        $\mathbf{D_a}$: dataset including weather-related accidents data for all the cities located in the considered area;
        $\mathbf{M}$: risk model
Result: $\mathbf{L}$: per-context risk levels
/* Context definition */
foreach $c \in \mathbf{C}$ do
    $\mathbf{R}_c \leftarrow$ partitionAreaIntoRegions($\mathbf{A}$, $c$)
    $\hat{\mathbf{R}}_c \leftarrow$: traceRiskyLocationsPerContext($\mathbf{R}_c$, $\mathbf{D_w}$, $\mathbf{D_a}$)
/* Weather data preparation */
$\mathbf{S} \leftarrow$ ExtractStations($\mathbf{D_w}$)
$\hat{\mathbf{S}} \leftarrow$: MapStationsToRiskyLocations($\hat{\mathbf{R}}_c$, $\mathbf{S}$)
foreach $s \in \mathbf{S}$ do
    $\mathbf{SE}_s \leftarrow$ ExtractAdverseEventSequences($\mathbf{D_w}$, $s$)
$\mathbf{SE} = \cup_{s \in \mathbf{S}} \mathbf{SE}_s$
/* Time series embedding */
$\mathbf{TSE} \leftarrow$ Train-Embeddings($\mathbf{SE}$, $\mathbf{S}$)
/* Apply the contextualized risk model */
foreach $c \in \mathbf{C}$ do
    $\mathbf{S}_c \leftarrow$ RetrievePertinentStations($\mathbf{S}$, $c$)
    $\hat{\mathbf{S}}_c \leftarrow$ PertinentStationsCloseToRiskyLocations($\mathbf{S}^c$, $\hat{\mathbf{S}}$)
    foreach $s \in \mathbf{S}_c$ do
        $r_s \leftarrow$: ApplyRiskModel($\mathbf{TSE}$, $\mathbf{M}$, $s$, $\hat{\mathbf{S}}_c$)
    $\mathbf{L}_c = \mathrm{Avg}_{s \in \mathbf{S}} \{r_s\}$
$\mathbf{L} = \cup_{c \in \mathbf{C}} \mathbf{L}_c$
return $\mathbf{L}$
```

A. SUMMARY OF THE NOTATION USED

- $\mathbf{A}$: geographical area under consideration.
- $\mathbf{R}$: delimited spatial region within $\mathbf{A}$.
- $\mathbf{S}$: set of meteorological stations.
- $\mathbf{W}$: set of weather elements (e.g., temperature, rain/snow level, snow depth) monitored by the stations in $\mathbf{S}$.
- $\mathbf{T}_s^w$: series of historical measurements of weather element $w \in \mathbf{W}$ acquired by station $s \in \mathbf{S}$.
- $\mathbf{F}$: set of features describing the key geographical and census properties of a region (e.g., *elevation, urbanization level*).
- $\mathbf{C}$: set of relevant contexts.
- $\mathbf{E}$: set of discrete events corresponding to pairs $\langle w, m \rangle$, where $w \in \mathbf{W}$ and m is a discrete measurement level for w.
- $\mathcal{P}(\mathbf{E})$: power set of $\mathbf{E}$.
- $\mathbf{SE}_s$: sequences of adverse weather events acquired by station $s \in \mathbf{S}$.
- $\hat{\mathbf{S}}$: subset of stations $\hat{s} \in \mathbf{S}$ labeled as *risky*.
- $\mathbf{L}_c$: contextual risk level.

B. CONTEXT DEFINITION

The geographical area $\mathbf{A}$ under consideration is partitioned into a set of regions $\mathbf{R}$. Each region is associated with a fixed set of features $\mathbf{F}$, which describe either geographical, cartographic, orographic, or census data. For example, the U.S. territories can be divided into regions characterized by different elevation range, district, country, and urbanization level.

Regions with similar characteristics are clustered into contextual groups. Specifically, a *context* c is defined as a subset of feature values (one or more). We assume that a set $\mathbf{C}$ of relevant contexts is provided as input by the domain expert. Set $\mathbf{R}_c$ groups all the regions characterized by context c. For example, a context may group all the regions in the U.S. characterized by a *high* urbanization level.

Given a collection of weather-related accident data $\mathbf{D_a}$, we enrich each context with the corresponding set of past road accident locations (see Line 3 in Algorithm 1). Specifically, for each context c we keep track of the GPS positions of all the road accidents occurred within $\mathbf{R}_c$. Tracking the geographical positions of past road accidents will allow us to map adverse weather conditions to risky locations, as discussed later on.

C. WEATHER DATA PREPARATION

Weather-related measurements are acquired by a set $\mathbf{S}$ of meteorological stations spread over the geographical area. Stations monitor various weather elements $\mathbf{W}$ (e.g., temperature, rain/snow level, snow depth).

For each context c we acquire the raw series of weather elements' measurements $\mathbf{T}_s^w$ reported by all the meteorological stations located within c and collect them into a unique *contextual weather dataset*. To define the risk level of a given context, each contextual dataset will be separately analyzed.

To study the influence of adverse weather effects on traffic safety, we first identify the occurrences of adverse weather conditions in the raw weather element measurements (see Lines 7-9 in Algorithm 1). To this aim, for each station s we extract the temporal sequence $\mathbf{SE}_s$ of discrete events reported by s that represent adverse weather conditions. More specifically, each event is a triple $\langle w, \Sigma, m\rangle$, where $w \in \mathbf{W}$ is a weather element (e.g., temperature), Σ is a comparison operator (e.g., $<$), and m is a threshold. For instance, event $\langle w = \text{temperature}, \Sigma = <, m = -20°C\rangle$ indicates the occurrence of a critical minimum temperature below $-20°$ C.

Event sequences $\mathbf{SE}_s$ will be considered to estimate the per-context risk levels as they embed all the weather-related risky patterns observed by station s, independently of the recorded occurrences of past road accidents.

D. TIME SERIES EMBEDDING

Time series embedding entails encoding the sequences of adverse weather events associated with all the input meteorological stations within the analyzed context (both located in risky locations and not) into a unified, high-dimensional vector space (see Line 11 in Algorithm 1). Each vector in the embedding space corresponds to a different station and embeds all the key information provided by the observed adverse weather events.

Figure 1 depicts the time series embedding inference process. Red and grey icons respectively correspond to stations located in risky locations and not. They are both mapped to vectors in the vector space according to the sequence of adverse weather events observed in the weather measurements they acquired.

To encode the sequences of adverse weather events we tailor the word-level embedding strategy called Paragraph2Vec [30], originally designed for text processing, to our context of analysis. The resulting vector space consists of a distributed vector representation of meteorological stations, where all the key information relative to each station is encoded into a separate vector. The key idea is to map stations reporting similar sequences of adverse weather events into the same region of the embedding space thus capturing the underlying characteristics of the seasonal trends in the event sequence. To this aim, we adopt an approach similar to a word-level text encoder, which focuses on capturing the semantic meaning of a word in a textual corpus.

Our purpose is to encode the words in a vocabulary $\mathcal{P}(\mathbf{E})$, where each word corresponds to one of the possible events combinations that may occur on a given day. To capture the seasonality of the adverse weather events, we split the historical sequences of adverse weather events observed by each station into multiple sub-sequences, each one corresponding to specific time spans (e.g., the yearly sequences). The aggregated sequences of daily event combinations observed by a station within a year virtually correspond to a text paragraph (consisting of a sequence of words).

The inference process of the paragraph-level embeddings is performed by an extended version of the Paragraph2Vec [31] architectures, namely the Distributed Memory-like (PV-DM-like) and Distributed Bag of Words (PV-DBOW-like), tailored to the problem under analysis. The PV-DM-like model, depicted in Figure 2(a), relies on the established Continuous Bag-of-Word (CBOW) model first proposed in Word2Vec [32]. According to the distributional hypothesis, CBOW infers the word vectors by assuming that the occurrences of a given word in a text are likely to be correlated with those of its immediately preceding or subsequent ones (namely the *context*). Analogously, here we predict the occurrence of specific adverse weather events on a given day based on its temporal correlation with the adverse weather events observed in the preceding/following days. The network takes as input the paragraph and station identifiers as well as the corresponding sequences of daily event combinations. It returns the encoding of the subsequent daily event combination.

PV-DBOW-like, depicted in Figure 2(b), disregards the contextual information at the input level and forecasts randomly sampled encodings of daily event combinations starting from either the paragraph ID or the station ID.

E. CONTEXTUALIZED RISK MODELS

This step entails estimating the risk level of each context on top of the time series embedding model (see Lines 12-19 in Algorithm 1). The returned risk level quantifies the likelihood

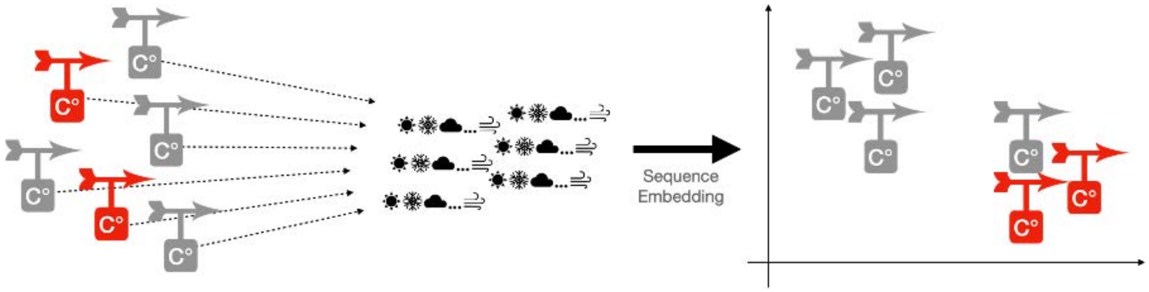
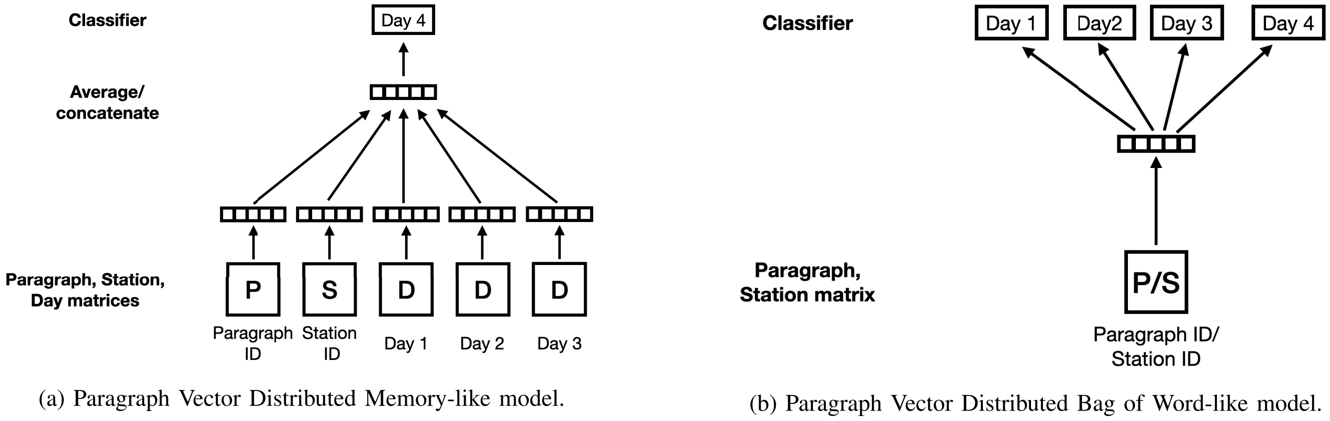


FIGURE 1. Inference of station embeddings based on adverse weather event series.



(a) Paragraph Vector Distributed Memory-like model.

(b) Paragraph Vector Distributed Bag of Word-like model.

FIGURE 2. The architectures.

that the presence of adverse weather effects influence road accidents occurrences within the analyzed context.

We present here four different strategies to compute the risk level tailored to the analyzed context, namely the *contextualized risk models*. The idea behind all the proposed risk models is to opportunistically reuse the (partial) information about the presence of risky locations to quantify how similar the embedding vectors of the stations located in non-risky locations are compared to the station vectors located in risky locations (within the same context). In a nutshell, the more similar the station embeddings the more risky the analyzed context because the weather element measurements are likely to be temporally correlated with one or more risky sequential patterns.

Whenever not otherwise specified, pairwise vector similarities are computed using the cosine similarity, which is known to be suitable for comparing samples in high-dimensional dataset [33].

To define the per-context risk levels, we designed the following models: (i) Centroid-based model (depicted in Figure 3a using a simplified bidimensional space), (ii) proximity-based model (see Figure 3b), (iii) density-based (see Figure 3c), and (iv) Top-k model (see Figure 3d). A thorough description of each model is given below.

1) CENTROID-BASED MODEL

Each station embedding belonging to the context is compared with the centroid of the stations located in the neighborhood of a risky location (within that context). The *centroid* is the representative station located at the center of the risky locations. It is computed as the point-wise average vector of all the station vectors located in a risky location. For example, the red cross in Figure 3a represents the centroid of the four stations located in risky locations (colored in red). To estimate the risk level of the light blue and light green stations we measure the respective distances from the centroid. The overall risk level of the context is computed by averaging the per-station risk levels.

2) PROXIMITY-BASED MODEL

Each station within the context is compared with the K nearest stations located in a risky location (where K is specified by the end-user). Similarities between pairs of stations are estimated using the pairwise vector distances. The risk level of the context is computed as the average of all the station levels.

For example, the light blue station vector in Figure 3b is compared with the 3 nearest station vectors corresponding to risky locations. Notice that the nearby grey stations are ignored as their corresponding risk levels are a priori unknown.

3) DENSITY-BASED MODEL

Similar to the homonym clustering technique [34], the neighborhood of each station vector is explored to quantify the *density* of stations located in risky locations. The neighborhood of a station is defined as the subset of stations whose similarity is above a given threshold t. For example, in Figure 3c the neighborhood of the light blue station (respectively light green) is represented by all the stations within the light blue circle (respectively light green). The risk level of a station is computed as the percentage ratio of risky stations to the overall number of neighbor stations. For example, the light blue station has just one station located in a risky location over three neighbor stations (i.e., one red station and two grey stations). Conversely, the light green station has ratio 1 as all the three neighbor stations are red. The overall risk level is the average risk level over all the stations within the context.

4) TOP-K MODEL

The Top-K model is analogous to the density-based one, but the concept of neighborhood is defined here as the top-K nearest stations (rather than all the stations that satisfy a minimum similarity threshold).

For example, the light blue and light green stations in Figure 3d are graphically linked to the corresponding top-3 nearest neighbors in the hyperspace. The corresponding risk levels are $\frac{1}{3}$ and 1, respectively.

IV. EXPERIMENTS

This section summarizes the main empirical results achieved on real-world dataset collecting past weather and road accidents data reported in the U.S.

The experiments were run on a machine equipped with Intel® Xeon® X5650, 32 GB of RAM and running Ubuntu 18.04.1 LTS.

The main settings used throughout the empirical analyses are summarized below.

- *Weather and traffic data*: we consider yearly sequences of adverse weather events annotated with road accidents occurrences at the daily granularity.
- *Time series embedding*: We infer the vector representation employing the Paragraph2Vec [31] model using the PV-DM (Distributed Memory Model of Paragraph Vectors) architecture, an embedding vector size 300, and 50 training epochs.
- *Weather incidence*: we set the risk threshold to the 75th percentile of the per-station Empirical Cumulative Distribution Function (ECDF).

The remainder of this section is organized as follows. Section IV-A and IV-B describe the analyzed data. Sections IV-C and IV-D respectively analyze the effect of the main input parameters of the designed method on the achieved results and compare the performance of the proposed risk models in different contexts. Finally, Section IV-F shows a prime example of graphical dashboards built on top of the computed risk levels.

A. DATASET DESCRIPTION

We employed a dataset integrating multiple data sources. Specifically, (i) a collection of time series describing the raw weather element measurements acquired by various meteorological stations located in the U.S., (ii) a collection of the road accidents that occurred from 2016 to 2020 in the U.S., and (iii) a selection of contextual pieces of information including cartographic, orographic, and census data related to the U.S. territories.

Hereafter we will describe the presented data sources in detail.

1) WEATHER DATA

We retrieved the daily values (since 1950) of weather measurements gathered by 1218 different meteorological control stations that are part of the U.S. Historical Climatology Network (USHCN). The aforesaid collection has been made available by the NOAA's National Centers for Environmental Information.[1] To avoid introducing a bias in the subsequent analyses, we will disregard those stations that are either placed far from all urban areas or have not enough historical data (i.e., the number of missing values is significant).

The retrieved dataset describes each station with multivariate time series. They consist of the daily gathered data samples about multiple weather elements. Specifically, for our purposes, we extracted the following information:

- Precipitation (PRCP)
- Snowfall (SNOW)
- Snow depth (SNWD)
- Maximum temperature (TMAX)
- Minimum temperature (TMIN)

It is worth noticing that, according to the data source documentation, the above-mentioned elements are referred to as the five *core elements* as they are reported for a larger portions of days compared to those of minor importance.

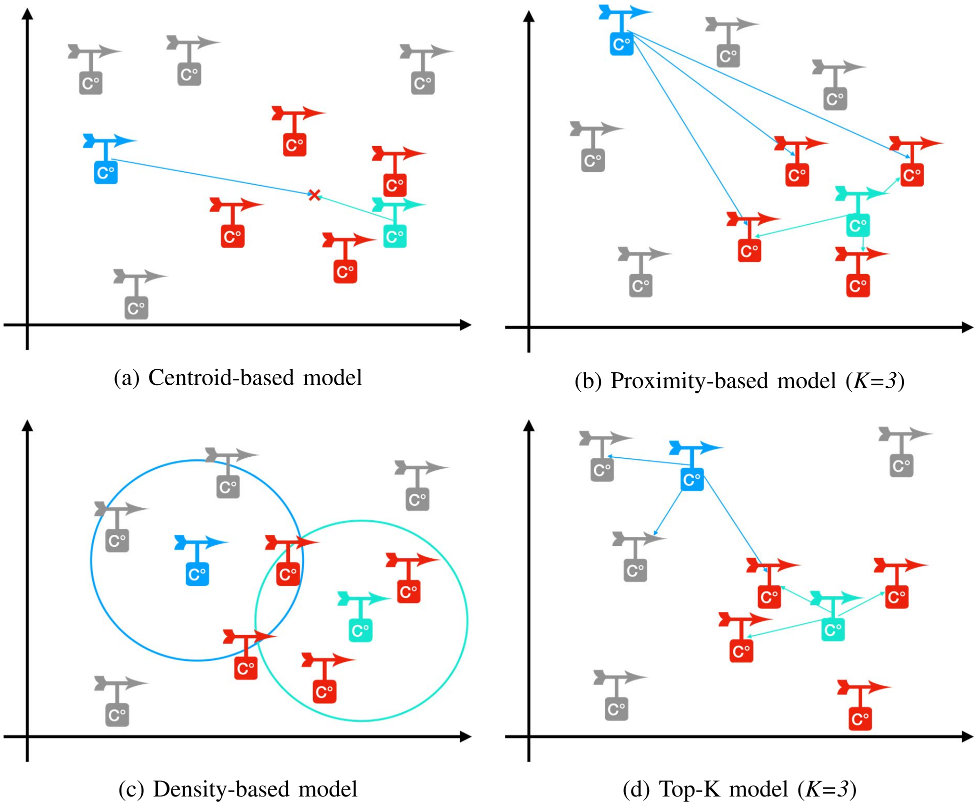
We extract the following *adverse weather events* from the complete dataset:

- Maximum temperature over 32° C
- Maximum temperature over 40° C
- Minimum temperature under -10° C
- Minimum temperature under -20° C
- Precipitation over 200mm
- Precipitation over 300mm
- Snowfall over 200mm
- Snowfall over 350mm
- Snow depth over 400mm
- Snow depth over 600mm

For our purposes, we also define a subset of more *severe weather events*. They consist of the five most restrictive events presented namely (i) Maximum temperature over 40° C, (ii) Minimum temperature under -20° C, (iii) Precipitation over 300mm, (iv) Snowfall over 350mm and (v) Snow depth over 600mm).

The thresholds defining *adverse* and *severe weather events* have been extracted from governmental sources[2],[3],[4],.[5]

[1] https://www.ncei.noaa.gov/ (last access: June 2021)
[2] https://www.weather.gov/ama/heatindex
[3] https://www.weather.gov/dlh/extremecold
[4] https://www.weather.gov/gsp/snow
[5] https://www.weather.gov/car/Warning_Criteria



(a) Centroid-based model

(b) Proximity-based model (*K=3*)

(c) Density-based model

(d) Top-K model (*K=3*)

FIGURE 3. Contextualized risk models.

The likelihood of occurrence of different events depends on the event type, the considered station, and the date. The event distribution is rather variable.

The analyzed dataset contains 52 different event combinations. More than half of them are relatively frequent, whereas all the remaining ones appear less than 100 times.

Within the time frame considered, the number of recorded events per station varies between 0 and 14624 (mean value: 4600 events per station). The analyzed distribution is quite balanced: most of the stations report an average number of events, with barely a dozen of stations reporting less than 100 events and less than 50 reporting less than 1000 events.

2) ROAD ACCIDENTS DATA

The dataset stores road accidents data [35], [36] corresponding to the road accidents that occurred in 49 U.S. states in a five-year period (i.e., from 2016 and 2020). The stored data include various features characterizing the accidents such as the location, time, date, and severity level of the accident.

To label cities (and stations) as risky or not, we introduce the concept of *weather incidence*. It indicates the percentage increase in the number of road accidents that occurred on days when any severe adverse weather event occurred compared to the accident count on the remaining days. Cities characterized by a weather incidence value over a given threshold are labeled as *risky* since they are likely to show an increase in the number and/or severity of accidents during days with adverse weather conditions. The risk level of the meteorological stations is, instead, related to their neighborhood. Specifically, all the stations located close to a risky city (i.e., the red-colored stations depicted in Figures 1 and 3a-3d) are labeled as *risky stations*.

The minimum threshold value used to define risky locations/stations has been empirically determined by plotting the Empirical Cumulative Distribution Function (ECDF) of the weather incidence values for all the considered stations (see Figure 4). Given a weather incidence value of x, the ECDF computes the percentage of stations associated with a weather incidence value lower or equal to x. In practice, 60% of the stations are not associated with any weather-related accident, whereas for about 15% of them the number of weather-related accidents at least doubled the number of accidents that happened with fair weather. To define the weather-incidence threshold value needed to discriminate between risky and non-risky stations/cities, we select the value corresponding to the 75th percentile of the empirical distribution (highlighted in orange in Figure 4).

3) CONTEXTUAL DATA

The dataset samples are annotated with various contextual features. Specifically, we gather (i) cartographic information about the division of the U.S. territories in regions, divisions, and states, (ii) orographic information elevation of the

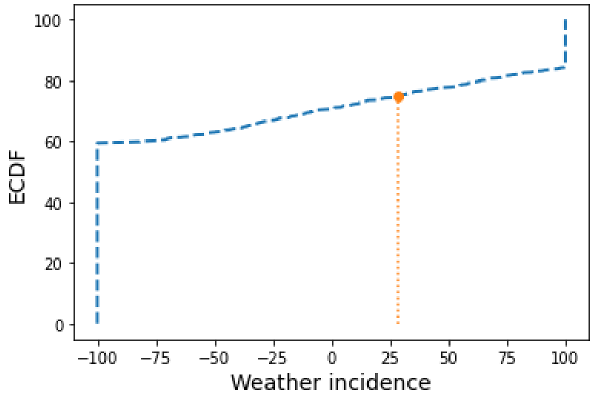


FIGURE 4. Empirical cumulative distribution function of the weather incidence values associated with all the meteorological stations.

location and (iii) census data indicating the urbanization level of a location.

Cartographic information is derived from the definitions of states, divisions, and regions defined by the U.S. Census,[6] whereas orographic information and urbanization levels about the main U.S. cities are extracted from a public dataset.[7]

a: CARTOGRAPHIC DATA

We characterize the location of each station based on the corresponding state, division, and region defined by the the U.S. Census.[8] The geographical distribution of stations over U.S. states is quite imbalanced, with a number of stations per state varying from 2 to 46. This implies that in several states we do not identify any risky station. However, as shown in Tables 1 and 2, the distribution of stations over divisions and regions is relatively even.

b: OROGRAPHIC DATA

Regarding the orographic properties of the weather stations considered, the elevation levels observed space from 59m under sea level, for a weather station in Death Valley (California), to 2763m above sea level for a station in Colorado. We discretize those values by grouping stations into elevation ranges presenting similar climates (see Table 3).

c: CENSUS DATA

To annotate the source data with census information, we first map every city to the nearest station. This results in each station having a list of cities under its influence. 13 U.S. cities are associated with each of the 969 stations that result mapped. The variability in the number of cities per station is significant: it ranges between 1 and 181. For this reason, to define the urbanization level of the neighborhood of a station, we set a cutoff threshold on the population size of a city under the station's influence. More specifically, employing the threshold presented in [37], we define the following three different ranges of population size:

- *High urbanization level*: the most populated city located in the station neighborhood has a population bigger than 250000 people.
- *Fair urbanization level*: the most populated city located in the station neighborhood has a population between 50000 and 250000 people.
- *Low urbanization level*: the most populated city located in the station neighborhood has a population smaller than 50000 people.

TABLE 1. Distribution of the meteorological stations over the urbanization levels.

Census region	Total num. of stations	Num. of risky stations	Percentage of risky stations
South	278	15	5.4
West	297	69	23.2
Northeast	98	34	34.7
Midwest	296	123	41.6

TABLE 2. Distribution of meteorological stations over the U.S. census divisions.

Census division	Total num. of stations	Num. of risky stations	Percentage of risky stations
East South Central	62	1	1.6
Mountain	186	58	31.2
West South Central	94	9	9.6
Pacific	111	11	9.9
New England	28	12	42.9
South Atlantic	122	5	4.1
East North Central	116	46	39.7
West North Central	180	77	42.8
Middle Atlantic	70	22	31.4

TABLE 3. Distribution of the meteorological stations over the elevation ranges.

Elevation range (Meters above sea level)	Total num. of stations	Num. of risky stations	Percentage of risky stations
[min, 400]	580	124	21.4
(400, 1000]	206	63	30.6
(1000, 1500]	96	22	22.9
(1500, max]	87	32	36.8

TABLE 4. Distribution of the meteorological stations over the urbanization levels.

Urbanization level	Total num. of stations	Num. of risky stations	Percentage of risky stations
High	139	29	20.9
Fair	206	54	26.2
Low	624	158	25.3

B. CHARACTERIZATION OF WEATHER-RELATED ACCIDENTS

We analyze here the distributions of accidents, weather-related and not, over each of the previously defined contextual features (see Tables 1-4).

Regarding the partitions into regions (see Table 1), the groups are quite well balanced. The total number of

[6] https://www.census.gov/ (latest access: November 2021)
[7] https://simplemaps.com/data/us-cities (latest access: June 2021)
[8] https://www.census.gov/ (latest access: June 2021)

TABLE 5. Weather-related road accidents counts per severity level.

U.S. Census region	Accidents count
South	1271745
West	1070892
Midwest	397477
Northeast	275834

TABLE 6. Weather-related road accidents counts per U.S. census division.

Census division	Accidents count
Pacific	889724
South Atlantic	711125
West South Central	436232
East North Central	287976
Middle Atlantic	237009
Mountain	181168
East South Central	124388
West North Central	109501
New England	38825

stations in the Northeast region is slightly lower because their extension is averagely higher. In the South region, the percentage of critical stations is quite low. This is consistent with the definition of critical station, which is correlated with the presence of extreme weather events such as heavy precipitations (e.g., rain, snow), cold waves or heat waves. Notice that in the South only the latter event type, and partially the rain precipitations, are widely present.

Concerning the Census division (see Table 2), similar observations hold. It is indeed clear that the divisions with the lowest percentage of critical stations are the divisions of the South region (West South Central, East South Central and South Atlantic) plus the Pacific division. Despite being on the opposite coast, Pacific division has a similar climate. As for the regions, in New England the total number of stations is low (28 stations), and this is again due to the limited size of the division (i.e., the smallest one in the analyzed data). In general, the distribution of the analyzed stations across the Census divisions is quite balanced. Tables 5 and 6 support the previous explanations of the distributions over census regions and divisions.

Focusing on the elevation ranges (see Table 3), it comes out that most of the stations are located under 400m of altitude. However, the locations at higher elevation ranges show a higher percentage of risky stations.

Regarding the urbanization levels (see Table 4), the resulting partitions are not perfectly balanced and this is probably due to the morphological properties of the United States of America. Nevertheless, the number of total and critical stations are sufficient for completing the target analyses and the percentage of critical stations is well balanced among the three different urbanization levels.

Finally, we also analyze the distribution of the analyzed data over the severity level of the road accident (see Table 7). As expected, most of the accidents report a medium severity level (i.e., levels 2 or 3).

TABLE 7. Weather-related road accidents counts per severity level.

Severity level	Accidents count
1	25423
2	2061835
3	838634
4	90056

C. CONFIGURATION SETTINGS FOR THE RISK MODELS

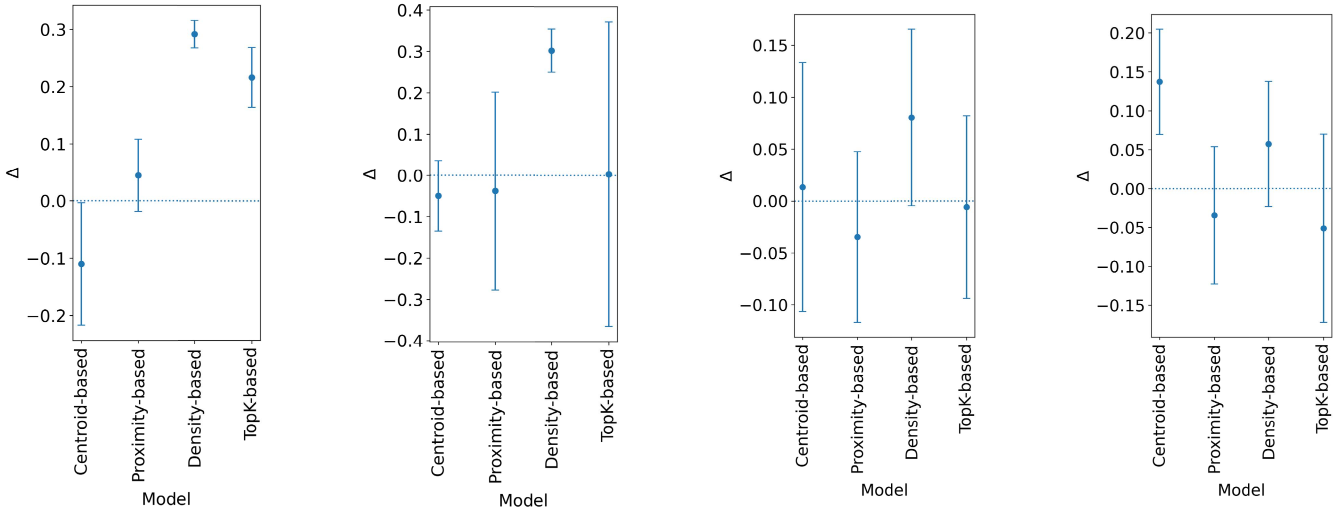
We focus here on the setup of the contextualized risk models. The density-based risk model requires the setting of a minimum similarity threshold (t), which is exploited to define the boundaries of a station neighborhood. All the stations that are highly similar to the reference station s in the vector space are considered s's neighbors. Similarly, to apply the top-K and proximity-based models the number K of most similar station vectors to be considered must be specified.

We perform a grid search to test various configuration settings by varying t in the range [0.6, 0.9], K for the top-K model between 150 and 1000, and K for the proximity model between 5 and 40. The recommended settings are $t = 0.75$ (Density-based), $K = 250$ (top-K), and $K = 15$ (Proximity-based). Notice that the recommended K values are relatively small. This indicates that including weakly similar stations in the risk model is likely to be potentially harmful.

D. PERFORMANCE COMPARISON BETWEEN RISK MODELS

We explored the applicability of the contextualized risk models to the real-life case study by applying the following steps:

1) Firstly, we computed the risk levels per context using the risk models presented in Section III-E. Risk levels are expected to reflect the incidence of adverse weather conditions on the road accidents occurrences within a specific context. The higher the risk levels, the higher the influence of weather-related events on traffic safety.
2) Secondly, separately for each model and context we identified the stations within each context that fall in the first and fourth quartiles according to contextualized risk level distribution. The latter stations are likely to have a higher risk of weather-related accident occurrences in the neighborhood. The former ones are expected to be the stations with the least risk level.
3) Thirdly, we verify the compliance of the risk model outcomes with the expected ground truth. The subset of stations belonging to the first quartile (according to the assignment described in the previous step) is expected to include a very limited number of risky stations, whereas the subset of stations in the fourth quartile is likely to include many risky stations.
4) Finally, we compared the outcomes of the contextual risk models to choose the best-performing ones in terms of maximal compliance with the ground truth.



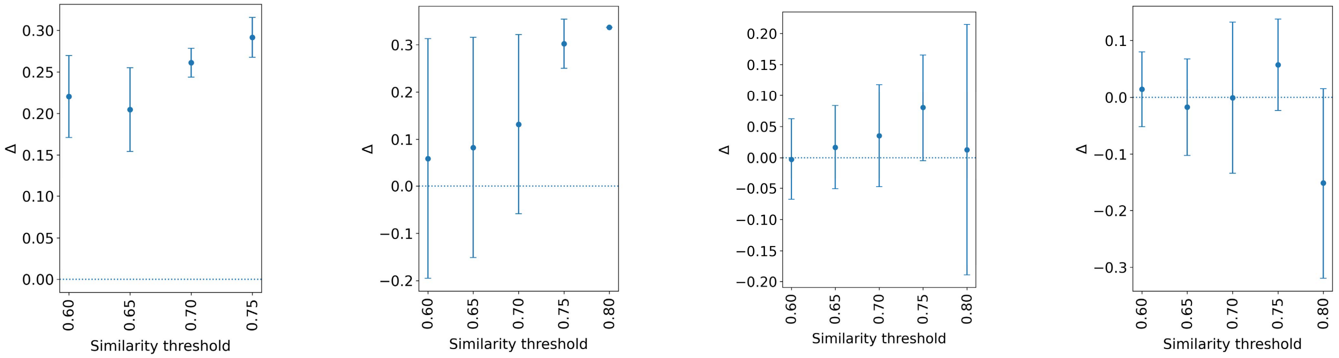
(a) Urbanization level

(b) Elevation range

(c) Census region

(d) Census division

FIGURE 5. Performance comparison between contextualized risk models against the Ground Truth.



(a) Urbanization level

(b) Elevation range

(c) Census region

(d) Census division

FIGURE 6. Comparison of different similarity threshold (*t*) values. Density-based risk model.

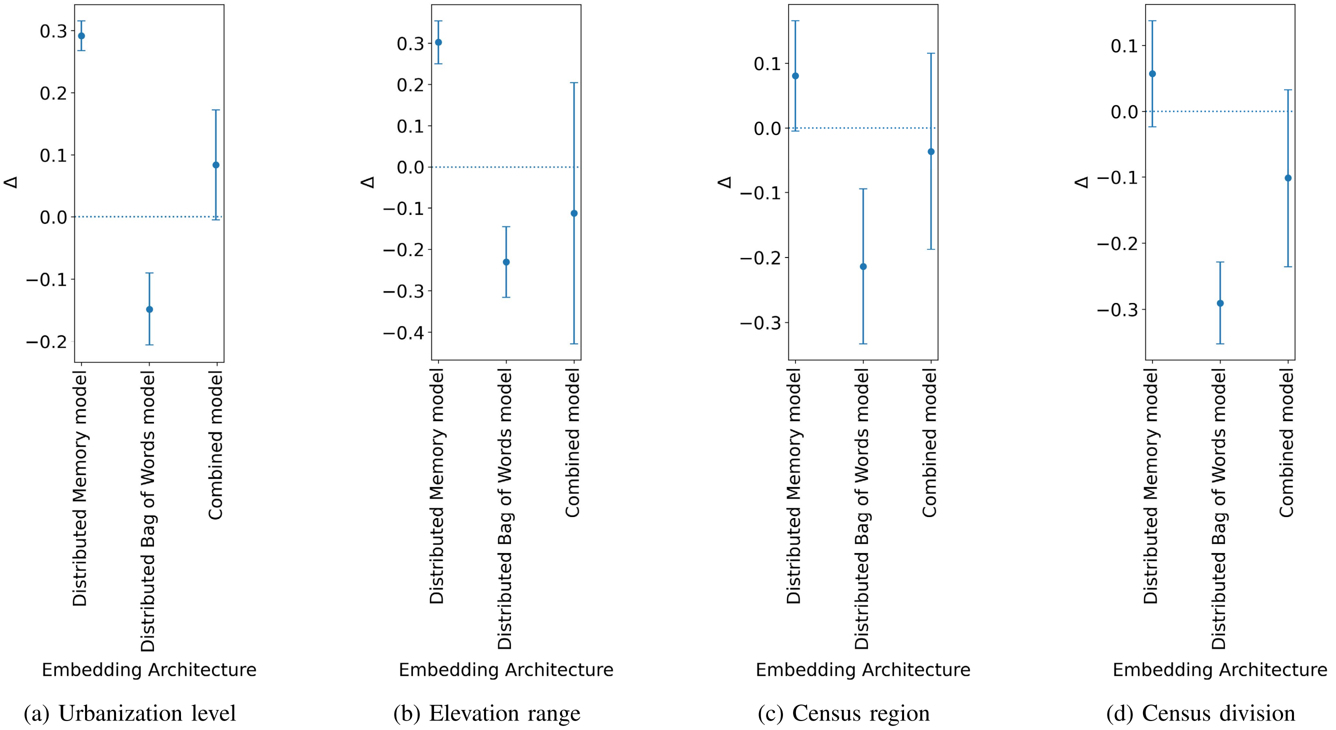
Figure 5 shows, separately for each context type, the performance of the proposed risk models. They are computed as the average and standard deviation of the differences between the ratio of the risky stations in the fourth quartile to the one in the first quartile (hereafter denoted by Δ). The average value is expected to be positive for all the analyzed contexts (i.e., the majority of the risky stations are in the fourth quartile, whereas only a few of them fall into the first quartile). The aforesaid indicator measures the ability of the model to properly assign the stations in the ground truth.

The density-based model was the only one that achieved consistently positive results over all the analyzed contexts, thus it was recommended as the reference risk model for the analyzed case study. For the sake of completeness, in Figure 6 we also report similar results achieved by the density-based model by varying the value of the similarity threshold t. The achieved results confirm that setting t to 0.75 is appropriate for the analyzed scenario.

E. PERFORMANCE COMPARISON BETWEEN EMBEDDING ARCHITECTURES

We compare here the performance of the two architectures used for time series embedding, i.e., PV-DM-like and PV-DBOW-like (see Section III-D). In compliance with [31], we also tested an ensemble method that combines PV-DM-like with PV-DBOW-like.

The results, summarized in Figure 7, show that the performance of PV-DBOW-like was always the worst, whereas that of PV-DM-like was averagely the best. Notice that, unlike PV-DBOW-like, PV-DM-like takes the seasonality of the adverse weather events into account while generating the paragraph-level encodings. This can be deemed as particularly helpful to capture the underlying correlations between weather event occurrences. Adopting an ensemble of PV-DM-like and PV-DBOW-like strategies did not yield significant improvements, likely because the achieved results are harmed by the poor PV-DBOW-like performance.



(a) Urbanization level (b) Elevation range (c) Census region (d) Census division

FIGURE 7. Comparison of different embedding architectures. Density-based risk model.

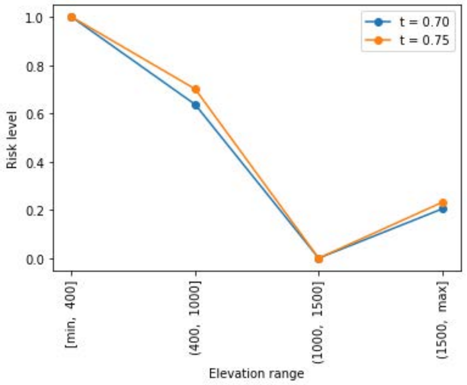


FIGURE 8. Risk levels for the contexts in the elevation value range. Strategy: Density-based.

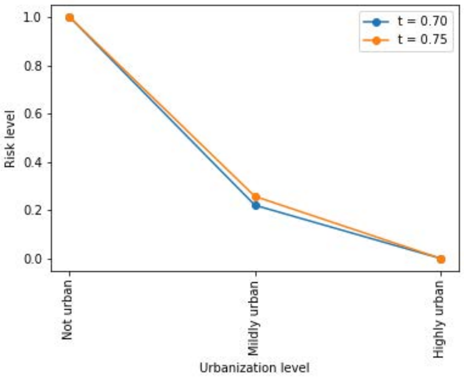


FIGURE 9. Risk levels for the contexts in the urbanisation level. Strategy: Density-based.

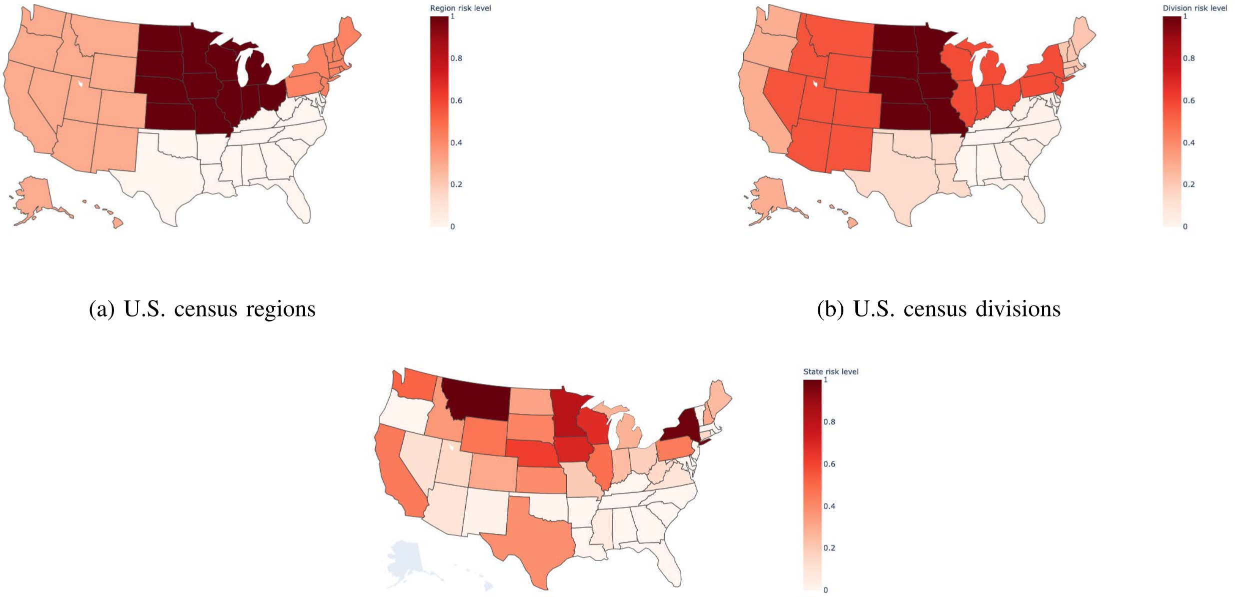
F. GRAPHICAL EXPLORATION OF PER-CONTEXT RISK LEVEL

We report here a prime example of the use of the per-context risk levels. Specifically, we compute the risk level of the various contexts separately for each attribute. Then, we average the KPIs over all the stations in the context. Depending on the attribute, we plot the risk levels using the most appropriate data representation (e.g., a map-based dashboard for the geographical features).

Let us consider the regions first (see Figure 10a). The South region turns out to be the one with the lowest probability of seeing an increment in road accidents during extreme weather days, whereas the Midwest is the region with the highest risk, followed by Northeast and West. This result is fully consistent with the official U.S. census data. The risk levels associated with the U.S. census divisions reflect the same distribution of the regions (see Figure 10b).

Two special cases are worth noticing: the *East North Central* (in the Midwest) and the *Middle Atlantic* (in the Northeast). These two divisions, which are geographically adjacent but belonging to different regions, show very similar risk levels (i.e., 0.579 and 0.577, respectively).

The distribution of the risk levels per state is quite similar to the previous ones (see Figure 10c). The only exception is the Texas state, which presents a risk level quite higher than its surroundings. The aforesaid preliminary finding is confirmed by the unexpectedly very low temperatures recorded in Texas during winter 2021.



(a) U.S. census regions

(b) U.S. census divisions

(c) U.S. states

FIGURE 10. Examples of map-based dashboards reporting the levels of risk of weather-related road accidents in different contexts. Density-based risk model. Similarity threshold $t = 0.75$.

TABLE 8. Execution times spent by the embedding models (expressed in seconds).

	Distributed Memory	Distributed Bag of Words	Combined DM-DBoW
Retrieving corpus	1.469		
Building vocabulary	24.833		
Training model	866.758	792.795	1659.553

TABLE 9. Execution times in seconds of the risk models.

Strategy	Execution time (in s)
Centroid-based	0.184 ± 0.02
Proximity-based	2.649 ± 0.03
Density-based	12.357 ± 0.07
TopK-based	7.764 ± 0.05

Focusing on the elevation ranges (see Figure 8), the risk level of the stations closer to ground level (i.e., stations with an elevation lower than 400m) has shown to be the highest one, followed by stations between 400m and 1000m. The least risky areas appear to be the ones in the elevation rage enclosed between 1000m and 1500m. The results achieved using different similarity threshold values (e.g., $t = 0.7$) do not show any significant variations.

Finally, the risk levels have shown to be inversely correlated to the urbanization level of an area. The risk of the least populated areas is higher than those of more urbanized ones. A possible explanation is that in urban areas citizens have easier access to alternative transportation means (e.g., train, subway).

G. EXECUTION TIMES

The overall computational times spent in each experimental session are in the order of thousands of seconds. The overall computation time is devoted to (i) data preparation, (ii) training of the embedding model and (iii) evaluation of the risk levels for each station in the analyzed context.

Data preparation, which took about 50% of the overall time (approximately 15 minutes), consists of the following sub-steps: (i) Transformation of the original dataset, (ii) Generation of the discrete events, (iii) Generation of the per-station event sequences. The execution time is roughly linear with the dataset cardinality.

Table 8 reports the execution times spent by the time series embedding methods. The (non-linear) representation learning process was, as expected, the most time-consuming stage of the presented methodology.

Finally, Table 9 indicates the time needed to evaluate the risk level, which is often negligible (around 12 seconds in the worst case). It slightly varies according to the analyzed context.

V. CONCLUSION AND FUTURE WORKS

The paper studied how to quantify the effect of extreme weather conditions on road accidents occurring in a specific spatial context, regardless of weather-unrelated influences. It relies on both a set of multivariate event series, encoded using a neural network-based embedding strategy, and an incomplete set of road accident annotations that characterize the stations located within the context. By conveniently reusing the partial information about risky locations the presented method is able to estimate a risk level per context.

Based on the experimental campaign conducted on real U.S. weather and traffic data we can conclude that

- The PV-DM architecture has shown to be the most suitable embedding model for tackling the problem under analysis.
- Adopting a density-based risk model appears to be the most effective strategy to capture the temporal correlations between adverse weather event sequences.
- Focusing on a relatively small number of neighbors (e.g., K=15) is, in general, more advisable to avoid degrading the overall risk model performance.
- Per-context risk levels can be effectively explored and analyzed through map-based dashboards.
- The estimated risk levels observed in a prime example meet the end-users' expectation to a large extent.

Future works will address (1) the integration of attention-based sequence encoders [38] to efficiently and effectively attend relevant temporal patterns in the event sequences, (2) the study and development of incremental, semi-supervised approaches that extend the currently proposed (static) methodology, and (3) the application of the proposed Machine Learning-based solution to different case studies (e.g., shared mobility services such as e-bikes and scooters, insurance customer profiling).

ACKNOWLEDGMENT

Computational resources were provided by HPC@POLITO, a project of academic computing within the Department of Control and Computer Engineering, Politecnico di Torino (http://www.hpc.polito.it).

REFERENCES

[1] A. Aloi, B. Alonso, J. Benavente, R. Cordera, E. Echániz, F. González, C. Ladisa, R. Lezama-Romanelli, Á. López-Parra, V. Mazzei, L. Perrucci, D. Prieto-Quintana, A. Rodríguez, and R. Sañudo, "Effects of the COVID-19 lockdown on urban mobility: Empirical evidence from the city of Santander (Spain)," *Sustainability*, vol. 12, no. 9, p. 3870, May 2020.

[2] A. K. Andersson and L. Chapman, "The impact of climate change on winter road maintenance and traffic accidents in West Midlands, UK," *Accident Anal. Prevention*, vol. 43, no. 1, pp. 284–289, Jan. 2011.

[3] B. Lenjani, P. Rashiti, D. Lenjani, P. Borovci, and N. Arslani, "Road accidents management and emergency medicine care," *Albanian J. Trauma Emergency Surg.*, vol. 3, no. 1, pp. 214–219, Jan. 2019.

[4] A. Theofilatos and G. Yannis, "A review of the effect of traffic and weather characteristics on road safety," *Accident Anal. Prevention*, vol. 72, pp. 244–256, Nov. 2014.

[5] H. Shin and J. Lee, "Temporal impulse of traffic accidents in South Korea," *IEEE Access*, vol. 8, pp. 38380–38390, 2020.

[6] J. Mihelj, A. Kos, and U. Sedlar, "Source reputation assessment in an IoT-based vehicular traffic monitoring system," in *Proc. Int. Conf. Identificat., Inf. Knowl. Internet Things, (IIKI)*, vol. 147, R. Bie, Y. Sun, and J. Yu, Eds. Beijing, China: Elsevier, 2018, pp. 295–299.

[7] M. T. Baldassarre, D. Caivano, D. Serrano, and E. Stroulia, "'Smart traffic': An IoT traffic monitoring system based on open source technologies on the cloud," in *Proc. 1st ACM SIGSOFT Int. Workshop Ensemble-Based Softw. Eng.*, A. Bucchiarone, M. Mongiello, F. Nocera, and M. Sheng, Eds., Lake Buena Vista, FL, USA, Nov. 2018 pp. 13–18.

[8] R. Jabbar, M. Shinoy, M. Kharbeche, K. Al-Khalifa, M. Krichen, and K. Barkaoui, "Urban traffic monitoring and modeling system: An IoT solution for enhancing road safety," in *Proc. Int. Conf. Internet Things, Embedded Syst. Commun. (IINTEC)*, Dec. 2019, pp. 13–18.

[9] A. I. Sunny, A. Zhao, L. Li, and S. K. K. Sakiliba, "Low-cost IoT-based sensor system: A case study on harsh environmental monitoring," *Sensors*, vol. 21, no. 1, p. 214, Dec. 2021.

[10] Z. Tao, "Advanced wavelet sampling algorithm for IoT based environmental monitoring and management," *Comput. Commun.*, vol. 150, pp. 547–555, Jan. 2020.

[11] V. Kishorebabu and R. Sravanthi, "Real time monitoring of environmental parameters using IoT," *Wireless Pers. Commun.*, vol. 112, no. 2, pp. 785–808, May 2020.

[12] M. J. Koetse and P. Rietveld, "The impact of climate change and weather on transport: An overview of empirical findings," *Transp. Res. D, Transp. Environ.*, vol. 14, no. 3, pp. 205–221, 2009.

[13] B. Deb, S. R. Khan, K. T. Hasan, A. H. Khan, and M. A. Alam, "Travel time prediction using machine learning and weather impact on traffic conditions," in *Proc. IEEE 5th Int. Conf. Converg. Technol. (I CT)*, Mar. 2019, pp. 1–8.

[14] R. Shang, Y. Zhang, Z.-J. M. Shen, and Y. Zhang, "Analyzing the effects of road type and rainy weather on fuel consumption and emissions: A mesoscopic model based on big traffic data," *IEEE Access*, vol. 9, pp. 62298–62315, 2021.

[15] R. Bergel-Hayat, M. Debbarh, C. Antoniou, and G. Yannis, "Explaining the road accident risk: Weather effects," *Accident Anal. Prevention*, vol. 60, pp. 456–465, Nov. 2013.

[16] T. Brijs, D. Karlis, and G. Wets, "Studying the effect of weather conditions on daily crash counts using a discrete time-series model," *Accident Anal. Prevention*, vol. 40, no. 3, pp. 1180–1190, May 2008.

[17] M. Schlögl, R. Stütz, G. Laaha, and M. Melcher, "A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset," *Accident Anal. Prevention*, vol. 127, pp. 134–149, Jun. 2019.

[18] M. Schlögl, "A multivariate analysis of environmental effects on road accident occurrence using a balanced bagging approach," *Accident Anal. Prevention*, vol. 136, Mar. 2020, Art. no. 105398.

[19] M. Schlögl and R. Stütz, "Methodological considerations with data uncertainty in road safety analysis," *Accident Anal. Prevention*, vol. 130, pp. 136–150, Sep. 2019.

[20] Y. Zou, Y. Zhang, and K. Cheng, "Exploring the impact of climate and extreme weather on fatal traffic accidents," *Sustainability*, vol. 13, no. 1, p. 390, Jan. 2021.

[21] Z.-Y. Zhan, Y.-M. Yu, T.-T. Chen, L.-J. Xu, and C.-Q. Ou, "Effects of hourly precipitation and temperature on road traffic casualties in Shenzhen, China (2010–2016): A time-stratified case-crossover study," *Sci. Total Environ.*, vol. 720, Jun. 2020, Art. no. 137482.

[22] R. Gallen, N. Hautière, A. Cord, and S. Glaser, "Supporting drivers in keeping safe speed in adverse weather conditions by mitigating the risk level," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 4, pp. 1558–1571, Dec. 2013.

[23] Y. He, M. Akin, Q. Yang, and X. Shi, "Conceptualizing how agencies could leverage weather-related connected vehicle application to enhance winter road services," *J. Cold Regions Eng.*, vol. 35, no. 3, Sep. 2021, Art. no. 04021011.

[24] M. E. Torbaghan, M. Sasidharan, L. Reardon, and L. C. W. Muchanga-Hvelplund, "Understanding the potential of emerging digital technologies for improving road safety," *Accident Anal. Prevention*, vol. 166, Mar. 2022, Art. no. 106543.

[25] H. M. Hassan and M. A. Abdel-Aty, "Predicting reduced visibility related crashes on freeways using real-time traffic flow data," *J. Saf. Res.*, vol. 45, pp. 29–36, Jun. 2013.

[26] D. Jaroszweski and T. McNamara, "The influence of rainfall on road accidents in urban areas: A weather radar approach," *Travel Behav. Soc.*, vol. 1, no. 1, pp. 15–21, Jan. 2014.

[27] F. Chen, S. Chen, and X. Ma, "Analysis of hourly crash likelihood using unbalanced panel data mixed logit model and real-time driving environmental big data," *J. Saf. Res.*, vol. 65, pp. 153–159, Jun. 2018.

[28] D. Kim, S. Jung, and S. Yoon, "Risk prediction for winter road accidents on expressways," *Appl. Sci.*, vol. 11, no. 20, p. 9534, Oct. 2021.

[29] J. Fior and L. Cagliero, "Estimating the incidence of adverse weather effects on road traffic safety using time series embeddings," in *Proc. IEEE 45th Annu. Comput., Softw., Appl. Conf. (COMPSAC)*, Madrid, Spain, Jul. 2021, pp. 13–18.

[30] C. Nalmpantis and D. Vrakas, "Signal2Vec: Time series embedding representation," in *Engineering Applications of Neural Networks* (Communications in Computer and Information Science), vol. 1000, J. MacIntyre, L. S. Iliadis, I. Maglogiannis, and C. Jayne, Eds. Crete, Greece: Springer, May 2019, pp. 80–90.

[31] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31st Int. Conf. Int. Conf. Mach. Learn. (ICML)*, vol. 32, 2014, pp. 1188–1196.

[32] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. 1st Int. Conf. Learn. Represent., (ICLR)*, Y. Bengio and Y. LeCun, Eds., Scottsdale, AZ, USA, May 2013, pp. 1–12.

[33] M. J. Zaki and W. Meira, *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2020.

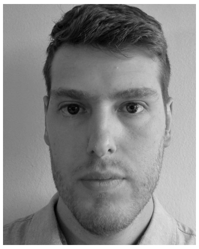
[34] P. Tan, M. S. Steinbach, A. Karpatne, and V. Kumar, *Introduction to Data Mining*, 2nd ed. London, U.K.: Pearson, 2019.

[35] S. Moosavi, M. H. Samavatian, S. Parthasarathy, and R. Ramnath, "A countrywide traffic accident dataset," 2019, *arXiv:1906.05409*.

[36] S. Moosavi, M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and R. Ramnath, "Accident risk prediction based on heterogeneous sparse data: New dataset and insights," 2019, *arXiv:1909.09638*.

[37] C. L. Ogden, C. D. Fryar, C. M. Hales, M. D. Carroll, Y. Aoki, and D. S. Freedman, "Differences in obesity prevalence by demographics and urbanization in US children and adolescents, 2013–2016," *JAMA*, vol. 319, pp. 2410–2418, Jun. 2018.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds. Long Beach, CA, USA, Dec. 2017, pp. 5998–6008.



JACOPO FIOR (Member, IEEE) received the bachelor's and master's degrees in computer science from the Università degli Studi di Torino (UniTO). He is currently pursuing the Ph.D. degree with the Department of Control and Computer Engineering, Politecnico di Torino. He has collaborated as a Research Assistant with the University of Helsinki. His current research interests include the study and application of machine learning and data mining techniques to time series data, and specifically to financial data.

LUCA CAGLIERO (Member, IEEE) received the master's degree in computer and communication networks and the Ph.D. degree in computer engineering from the Politecnico di Torino. He has been an Associate Professor with the Dipartimento di Automatica e Informatica, Politecnico di Torino, since January 2020. He teaches B.Sc., master-level, and Ph.D. courses on database systems and data mining techniques. His current research interests include the fields of machine learning, text mining, and deep NLP. Specifically, he has worked on text summarization, classification, and association rule mining. He has coauthored more than 100 scientific articles, including more than 40 international journals. He is currently an Associate Editor of the *ESWA* (Elsevier) and *MLWA* (Elsevier) journals. He has been the contact person of various consulting and research contracts. He was a recipient of the Working Capital Research Grant 2012 for the research project on web document summarization.