**Accident Forecasting in Saint Petersburg: A DBSCAN-Based Approach to Hotspot Detection and Temporal Analysis**

**Abstract**

Traffic accidents remain a critical challenge for urban safety management, with significant social and economic consequences. This study analyzes traffic accident data from Saint Petersburg (2022–2024) using advanced spatial and temporal methodologies to identify accident hotspots and explore patterns of occurrence. By employing the DBSCAN clustering algorithm, the study highlights dense accident zones where interventions are most needed. However, efforts to predict accident dynamics using temporal and weather-related features revealed limited predictive power, attributed to the sparse and stochastic nature of the data. The findings underline the importance of data granularity and hotspot-based analysis for effective traffic safety strategies.

**Introduction**

Traffic accidents are among the leading causes of mortality and injury in urban settings, necessitating targeted interventions for prevention and control. As cities grow, the complexity of traffic systems and road user behavior increases, making it vital to use data-driven methods for safety planning. Traditional approaches, such as using administrative boundaries to allocate resources, often fail to capture the complexity of real-world accident dynamics. Consequently, there is a growing demand for analytical techniques that consider spatial and temporal patterns to identify high-risk zones or "hotspots."

This study investigates accident trends in Saint Petersburg using traffic data from January 2022 to September 2024. The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm was chosen to identify accident hotspots because of its ability to detect clusters of arbitrary shapes and handle noise points. The results of this analysis were supplemented by an exploration of intra-day and intra-week seasonality and an evaluation of the determinants of accidents using decision trees and classification models. While the clustering analysis provided meaningful insights, predictive modeling revealed limited success, highlighting key challenges in accident forecasting.

The research addresses the following key objectives:

1. To identify high-risk accident hotspots in Saint Petersburg using spatial clustering.
2. To analyze temporal patterns of accidents to reveal seasonality and trends.
3. To assess the predictive power of weather, temporal features, and lagged accident data in modeling accident likelihood.

**Methodology**

**Data Description and Preprocessing**

The dataset consisted of 7,235 traffic accident records in Saint Petersburg over a 33-month period. Each record included geographic coordinates (latitude and longitude), timestamps, severity classification (“Light,” “Severe,” or “Fatal”), and regional identifiers. The following preprocessing steps were performed to ensure data quality and consistency:

1. **Outlier Removal:** Geospatial outliers were excluded based on pre-defined geographic boundaries to ensure only accidents within Saint Petersburg were analyzed.
2. **Translation and Standardization:** Regional names were transliterated from Russian to English, and severity classifications were standardized. For analytical clarity, accidents were grouped into two categories: "Light" and "Severe/Fatal."
3. **Temporal Filtering:** Only accidents that occurred between 2022 and 2024 with valid coordinates were retained.

**Spatial Clustering Using DBSCAN**

DBSCAN was applied to identify accident hotspots. The algorithm requires two key parameters:

* **Epsilon (ε):** Specifies the maximum distance between points to form a cluster. A value of 0.01 (~1 km) was chosen based on the scale of urban infrastructure.
* **MinPts:** Defines the minimum number of points required to form a cluster. A value of 100 ensured that clusters represented meaningful accident densities.

The algorithm grouped accidents into clusters of high density, assigning isolated incidents as noise. These clusters were visualized on a map, providing a clear representation of high-risk zones. Clusters corresponding to intersections, high-traffic areas, and poorly designed road segments were noted as critical areas for intervention.

**Temporal Dynamics and Seasonality Analysis**

Accident data was analyzed to explore patterns within weeks and days:

* **Day of the Week:** Accidents were grouped by weekday, revealing differences between workdays and weekends.
* **Hour of the Day:** Temporal distribution across 24 hours was assessed to detect peak times of occurrence.

This analysis aimed to identify trends that could guide policy measures, such as traffic regulation and public awareness campaigns.

**Predictive Modeling of Accident Dynamics**

Two complementary modeling approaches were employed to assess determinants of accidents:

1. **Regression Tree Models:** These models predicted accident counts for each administrative district and later for DBSCAN clusters. Predictors included weather data (e.g., temperature, precipitation, wind speed), temporal variables (e.g., hour of the day, day of the week), and lagged accident counts.
2. **Classification Models:** Binary classification models were used to predict the probability of accident occurrence within each DBSCAN cluster, leveraging the same set of predictors.

Lagged variables were introduced to capture temporal dependencies, and weather data was appended to account for external conditions.

**Results**

**Hotspot Identification with DBSCAN**

The DBSCAN algorithm identified several high-density clusters across Saint Petersburg, with notable concentrations at major intersections, roundabouts, and arterial roads. For instance, hotspots in the Nevskij and Primorskij districts corresponded to known traffic bottlenecks. These clusters provide actionable insights for targeted interventions, such as redesigning intersections, improving signage, or implementing speed control measures.

**Temporal Patterns**

Analysis of intra-day and intra-week patterns showed:

* Slightly higher accident frequencies on weekdays compared to weekends, likely due to increased commuter traffic.
* Peak accident times during morning (8–10 AM) and evening (5–7 PM) rush hours, consistent with urban commuting patterns.

However, these patterns were not pronounced, reflecting a relatively uniform distribution of accidents throughout the week.

**Predictive Modeling Outcomes**

Predictive models yielded limited results:

1. **Regression Trees:** Accident counts were poorly predicted, with most models returning trivial rules, such as "accidents are less likely before 8 AM." Even these insights were limited to specific clusters.
2. **Classification Models:** Binary classification models achieved low accuracy. Sparse events within clusters and inconsistent patterns hindered model performance.

The analysis suggests that the stochastic nature of traffic accidents, combined with sparse data, limits the predictive power of weather and temporal features.

**Discussion**

**Challenges in Predictive Modeling**

The limited success of predictive models underscores inherent challenges in accident forecasting:

* **Data Sparsity:** Many DBSCAN clusters contained only a few accidents over the study period, reducing the statistical power of models.
* **Complex Causality:** Traffic accidents are influenced by a wide array of factors, including driver behavior, road conditions, and external events, which were not captured in the dataset.
* **Cluster Resolution:** While DBSCAN improved spatial granularity compared to administrative boundaries, further refinement (e.g., dynamic clustering) may yield better results.

**Significance of DBSCAN Clustering**

Despite limitations in predictive modeling, DBSCAN proved effective in identifying meaningful accident hotspots. The ability to exclude noise points and adapt to the spatial distribution of data makes it a valuable tool for urban safety planning. Hotspot-based analysis allows for focused interventions, reducing the risk of accidents in high-density zones.

**Speculative Improvements**

To address current limitations, future work could explore:

* **Enhanced Data Collection:** Incorporate traffic flow, road quality, and driver demographics.
* **Dynamic Clustering:** Adjust DBSCAN parameters based on temporal variations to capture seasonal shifts in accident patterns.
* **Machine Learning Models:** Experiment with advanced models (e.g., neural networks) that can better handle sparse and stochastic data.

**Conclusion**

This study demonstrates the utility of DBSCAN for identifying traffic accident hotspots in Saint Petersburg. While predictive modeling revealed limited insights, the clustering approach highlights critical zones for safety interventions. The findings emphasize the need for more granular data and advanced analytical methods to improve accident forecasting. Future research should integrate richer datasets and adaptive clustering techniques to address the challenges identified in this study.