August 2024

Operationalized Social Media and Linear Directional Mean Address Drug Overdose Crisis in America

Anthony Corso
California Baptist University, acorso@calbaptist.edu

Esther Lee
California Baptist University, elee@calbaptist.edu

Nathan A. Corso
California Baptist University, nathan.corso@calbaptist.edu

Benjamin Sanders
California Baptist University, bsanders@calbaptist.edu

Operationalized Social Media and Linear Directional Mean Address Drug Overdose Crisis in America

Emergent Research Forum (ERF)

Anthony J. Corso
California Baptist University
acorso@calbaptist.edu

Esther Lee
California Baptist University
elee@calbaptist.edu

Nathan A. Corso
California Baptist University
nathan.corso@calbaptist.edu

Benjamin Sanders
California Baptist University
bsanders@calbaptist.edu

Abstract

This study examines social media's spatial distribution of cognition and drug-related language by employing the linear directional mean (LDM) and circular variance to analyze directional data. The results reveal differences in spatial distributions, with drug-referenced social media exhibiting a slightly stronger non-uniform pattern than cognition. Researchers should interpret findings cautiously due to several limitations, including temporal constraints, social media origin uncertainties, and feature extraction methods. The study introduces a novel artifact, demonstrating the effectiveness of enhanced grammar-based feature engineering. Additionally, natural language processing (NLP) extracts valuable insights from sparse text social media. Finally, this work extends current research with a deeper understanding of social media data utilization for various applications and underscores the potential of structured frameworks and advanced NLP approaches in this evolving domain.

Keywords

Social Media, Linear Directional Mean, Natural Language Processing.

Introduction

Decision Support Systems (DSS) are crucial in various domains, ranging from urban planning to predictive policing and transportation systems. These systems serve diverse functions, including enhancing safety, ensuring service continuity, minimizing costs, and other critical goals (Keenan and Jankowski 2019) (Pick et al. 2017). The increasing availability of social media data, big data infrastructure, and powerful computers fuel the development of data-driven (DD) algorithms to tackle complex societal challenges. Data-driven approaches are well-suited for scenarios with limited understanding of the underlying systems or highly complex and uncertain relationships between inputs, processes, and outputs (Smallman and Rieth 2017). Furthermore, machine learning and deep learning offer promising capabilities for such challenges.

Spatial Decision Support Systems (SDSS) represent a growing subfield within DSS research. This field explores the intersection of traditional statistical techniques, which often assume normal distributions and linear relationships, with the complexities of circular data and spatial phenomena. Advancements in Geographic Information Systems (GIS) and machine learning solutions are key facilitators of this convergence, making SDSS well-suited to support future applications (Afyouni et al., 2022).

This project investigates the potential of using social media data to improve understanding of spatial patterns in drug overdose occurrences. A novel approach uses the modified normalized difference in Linear Directional Mean (LDM) to analyze changes in two operationalized social media features across one location over nine years. The primary goals are twofold: first, to assess the efficacy of LDM in identifying trends and changes in drug overdose hot spots by leveraging statistically significant social media variables related to the crisis. Second, to explore the potential of LDM for analyzing trends associated with linear features, particularly within the context of social media data. Additionally, the project investigates the seamless integration of various data representations (graphical and tabular) within a spatial Decision Support System (SDSS) framework. Based on the context provided, the hypothesis is as follows:

H_o: Null Hypothesis: The LDM of cognition is equal to the LDM of drug-referenced social media.

H_a: Alternative Hypothesis: The LDM of cognition is not equal to the LDM of drug-referenced social media.

The rest of this study follows a structured progression, beginning with a comprehensive literature review and a detailed DSR research approach integrating longitudinal data collection and H3 spatial indexing. Subsequent sections explain the sources and collection methods for relevant data points, analyze spatial distribution patterns, and interpret findings in the context of existing literature. The study concludes with a comprehensive summary of key findings, implications, and recommendations for future research.

Literature Review

This literature review investigates the potential of longitudinal spatial analysis techniques, encompassing spatial autocorrelation, cluster detection, and H3 spatial indexing, to inform decision-making across diverse domains. These techniques offer valuable insights into temporal dynamics, spatial patterns, and directional trends, facilitating informed decision-making and predictive modeling in various fields. A study by Chainey et al., (2008) highlights the significance of incorporating temporal trends into spatial analysis. Longitudinal data collection methods enable researchers to track changes in spatial distribution patterns over time, revealing valuable insights into seasonal and long-term trends as demonstrated by Jones et al., (2012) in their longitudinal analysis of air temperature variations. Additionally, Afyouni et al., (2022) present a comprehensive survey of techniques for social event detection utilizing multiple features, modalities, and data sources, emphasizing the potential of integrating heterogeneous data sources to improve the accuracy and robustness of such systems. Spatial autocorrelation and cluster detection methods are crucial in identifying spatial patterns and clusters within longitudinal data. Research by Diggle et al., (2013) demonstrates the application of these techniques to detect spatial dependence and clustering of crime incidents across multiple time points, facilitating targeted interventions and resource allocation in high-risk areas. Recent studies have explored H3 spatial indexing in longitudinal cross-sectional spatial analysis, offering a granular approach to spatial data representation. Li et al. (2020) applied H3 to optimize urban spatial structure based on taxi trajectory data, while Tigges et al., (2023) utilized H3 to analyze spatial sentiment patterns in large language models, demonstrating its potential for diverse applications.

Research Approach

This study adopts Design Science Research (DSR) methodology as the overarching framework for the entire research project. The research approach integrates classical experimental design principles to explore the potential relationship between social media data, operationalized for author cognition and drug references. Following the established guidelines of a true field experiment (Bryman 2012), a specific type of classical design to rigorously test the hypothesis outlined in the introduction (Leroy 2011) is implemented.

The collection of social media data relied on naturalistic observation of tweet authors within the Seattle, Washington area. The dataset covers the period from September 2014 to March 2023, comprising approximately 43,960,000 tweets obtained via the Twitter stream. Figure 1 illustrates the five specific years analyzed in this project and the corresponding number of tweets utilized from each year.

2014 | 171,760	2016 | 107,220	2018 | 66,280	2020 | 32,760	2022 | 22,560

Figure 1. Tweet Data Collection in Seattle, Washington (2014-2023)

Each individual tweet, considered as the unit of analysis, underwent natural language processing techniques, and was subsequently tagged with the measures outlined in Table 1. For example, a tweet exhibiting a certain level of emotional perception related to drugs or its associated risks was labeled as

"YesCog" or "NoCog," depending on whether the author observed an event. Similarly, a tweet was measured for the presence of drug-related content through emojis or text and labeled as "DrugYes" or "DrugNo."

High-level Social Media Concept	Description	Levels of Measure [Source Code Variables]
Cognition	The variable will quantify the extent of emotional perception shaped by the cognitive processes of observation real-world events and their associated risks.	Yes (perceived event) No (no perceived event) [YesCog, NoCog]
Drug	This variable measures the subjective response or feeling expressed in a piece of text and quantifies the frequency and manner in which drugs are mentioned or discussed on social media platforms.	DrugYes, DrugNo [DrugPresentEmoji, DrugPresentText, DrugPresentTotal]

Table 1. Measures of NLP-Processed Social Media

Figure 2 illustrates the H3 spatial index utilized in this study, operating at a resolution level of eight. This level of resolution signifies the degree of detail in the spatial partitioning. Each tweet was spatially joined with the H3 layer; the visualization of social media activity for 2014 overlayed on it is depicted in Figure 3. With each tweet linked to its corresponding hexagonal cell (H3 cell) tweets could then be analyzed according to their geographic origins within these hexagonal cells.

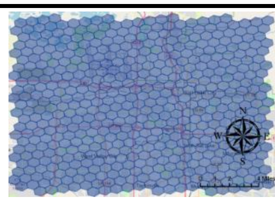


Figure 2. H3 Indexing

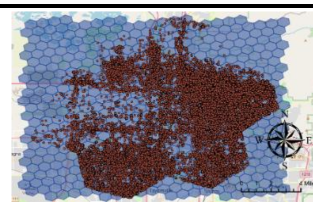


Figure 3. 2014 Tweets Joined with H3

An optimized hot spot analysis was conducted to examine significant clusters related to the spatial distribution of both the cognition and drug variables within the geographic area of interest. Figure 4a (cognition) and 4d (drug) depict the most significant hotspot for each variable for one year. Clusters highlighted in red (hotspots), characterized by high values surrounded by other high values, indicate areas of high activity of occurrences related to the cognition (4a) and drug (4d) variables. Conversely, blue clusters (coldspots) represent areas of low activity or occurrences. Figures 4b and 4e show the centroid of each hotspot cluster shown in 4a and 4b. The same process was repeated for all years and their respective polylines (lines connecting the centroids) were created to visually represent the spatial distribution of each cluster. Figure 4c and 4f show the polyline connecting each centroid, for each year, across the study area.

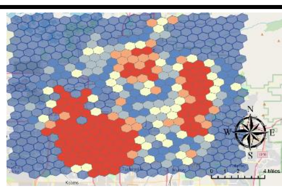


Figure 4a. Significant Cluster

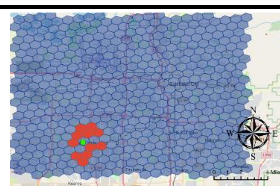


Figure 4b. Center of Cluster

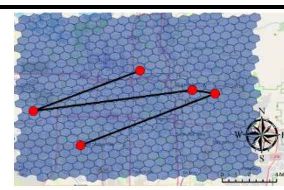


Figure 4c. Polyline

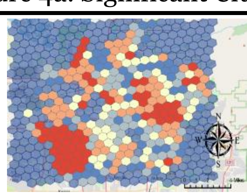


Figure 4d. Significant Cluster

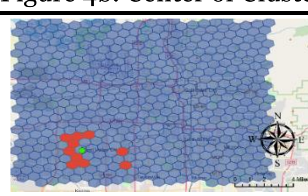


Figure 4e. Center of Cluster

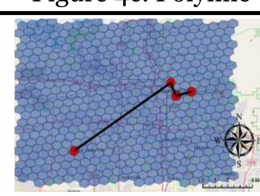


Figure 4f. Polyline

Figure 4. Hotspots and Centrids

The linear directional mean was calculated based on the polyline shown in Figure 4c and 4f and results are shown in Figure 5. This longitudinal approach enables the examination of trends and variations in spatial distribution patterns over time, enhancing the robustness of the findings.

Results

Analyzing the relationship between the cognition and drug variables provides valuable insight into the potential of social media data for informing and supporting drug overdose prevention efforts. By leveraging DSR and a true experiment this project defines a causal relationship between social media content (independent variable) and the observed changes in drug referencing (dependent variable), providing more substantial evidence than a simple correlation study. The study employed the linear directional mean (LDM) model to investigate the average direction of hotspot changes, focusing on how the predominant direction of these hotspots shifted or remained stable throughout the study period. The primary results revolve around the findings of the uniformity test, which is crucial for determining whether the set of angles or directions representing the hotspots is uniformly distributed around the circle or if there is a direction.

The null hypothesis (H0) states that the spatial distribution of cognition direction is equal to the drug direction of statistically significant optimized hot spot centroids based on an H3 spatial index resolution level of eight and follows a uniform or random pattern. The alternative hypothesis (HA) suggests that the spatial distribution of cognition direction is not equal to the drug direction of statistically significant optimized hotspot centroids based on an H3 spatial index resolution level of eight and follows a uniform or

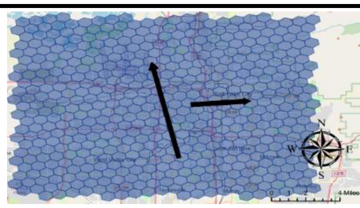
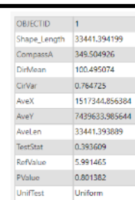


Figure 5a. LDM Vectors Cognition Left, Drug Right



OBJECTID	1
Shape_Length	33441.394199
CompassA	349.504926
DirMean	100.495074
CirVar	0.764725
AveX	1517344.856384
AveY	7439633.985644
AveLen	33441.393889
TestStat	0.393609
RefValue	5.991465
PValue	0.801382
UnifTest	Uniform

Figure 5b. Cognition Vector

OBJECTID	1
Shape_Length	12864.743667
CompassA	84.896317
DirMean	5.103683
CirVar	0.249459
AveX	1529607.640545
AveY	7442483.400015
AveLen	12864.743634
TestStat	4.577828
RefValue	5.991465
PValue	0.105057
UnifTest	Uniform

Figure 5c. Drug Vector

Figure 5. LDM Results

random pattern. Figure 5c shows that the drug vector is statistically significant at .105. In addition, the UnifTest is equal to Uniform for both vectors so the outcome of the hypothesis test is: fail to accept the null.

Discussion, Limitations, and Future Research

Discussion

The analysis reveals distinct spatial patterns for both Drug and Cognition variables based on the Linear Directional Mean (LDM) metric. The Drug variable exhibits a mean direction (DirMean) of 100.495 with a circular variance (CirVar) of 0.765, indicating a slightly stronger directional trend compared to the Cognition variable with a mean direction of 5.104 and a circular variance of 0.249. While both p-values of the uniformity tests exceed the significance level (0.05), suggesting the patterns are not statistically different from a uniform distribution, the Drug variable's p-value (0.801) is closer to 1, hinting at stronger evidence against randomness compared to Cognition (0.105). These findings imply that the observed hotspots for the drug directions are not uniformly distributed around the compass circle. It is likely that there is a prevailing drug direction or clustering of drug-related patterns. Findings also suggest potential spatial variations in sentiment and cognitive aspects associated with drugs and cognition on social media, warranting further investigation with additional spatial analysis techniques to strengthen understanding of these complexities.

Limitations

The research faced several limitations. Firstly, temporal constraints on data availability hindered the capture of short-term spatial changes, relying heavily on long-term trends. Secondly, using the standard REST API instead of the Firehose API resulted in a smaller data corpus, potentially introducing sampling bias. Thirdly, uncertainties about tweet origins due to user location settings affect spatial patterns. Fourthly, feature extraction methods like tokenization and part-of-speech tagging may have influenced results due to sparse tweet content. Lastly, while the linear directional mean and circular variance provided useful summaries, they may have oversimplified spatial patterns. Additionally, the chosen resolution level of the H3 spatial index and assumptions of data independence could have influenced outcomes, highlighting the need for refining methodologies in future research to enhance accuracy and reliability.

Future Research

Future research should address limitations in using LDM to understand the intricate interplay between sentiment and cognition on social media. This could involve expanding beyond LDM to capture richer spatial patterns; investigating approaches that relax assumptions and account for non-uniformity dependencies in the data; and enriching LDM analysis by combining it with advanced NLP techniques. This integration could include tasks such as fine-grained grammatical tagging, nuanced sentiment analysis, and contextual exploration, providing deeper insights into social media dynamics.

Conclusion

This study explored the spatial distribution of cognition and drug-related language on social media, utilizing the linear directional mean (LDM) and circular variance for analyzing directional data. While the results suggest differences in their spatial distributions, with drugs potentially displaying a slightly stronger non-uniform pattern compared to cognition, these findings warrant cautious interpretation due to several limitations. Moreover, the study introduces a novel artifact, demonstrating the effectiveness of enhanced grammar-based, feature-engineered NLP techniques in extracting valuable insights from social media communication, especially within the context of sparse text data. This research represents a significant advancement in understanding and harnessing social media data for various applications, highlighting the potential of structured frameworks and advanced NLP approaches in this evolving domain.

References

Afyouni, I., Aghbari, Z. A., and Razack, R. A. 2022. "Multi-Feature, Multi-Modal, and Multi-Source Social Event Detection: A Comprehensive Survey," *Information Fusion* (79), pp. 279–308. (https://doi.org/10.1016/j.inffus.2021.10.013).

Bryman, A. 2012. *Social Research Methods*, (4th ed.), Oxford ; New York: Oxford University Press.

Chainey, S., Tompson, L., and Uhlig, S. 2008. "The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime," *Security Journal* (21:1–2), pp. 4–28. (https://doi.org/10.1057/palgrave.sj.8350066).

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. 2013. "Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm," *Statistical Science* (28:4). (https://doi.org/10.1214/13-STS441).

Jones, P. D., Lister, D. H., Osborn, T. J., Harpham, C., Salmon, M., and Morice, C. P. 2012. "Hemispheric and Large-Scale Land-Surface Air Temperature Variations: An Extensive Revision and an Update to 2010," *Journal of Geophysical Research: Atmospheres* (117:D5). (https://doi.org/10.1029/2011JD017139).

Keenan, P. B., and Jankowski, P. 2019. "Spatial Decision Support Systems: Three Decades On," *Decision Support Systems* (116), pp. 64–76. (https://doi.org/10.1016/j.dss.2018.10.010).

Leroy, G. 2011. *Designing User Studies in Informatics*, Health Informatics, London: Springer London. (https://doi.org/10.1007/978-0-85729-622-1).

Li, M, and Emmanuel Stefanakis. "Geospatial operations of discrete global grid systems—A comparison with traditional GIS." Journal of Geovisualization and Spatial Analysis 4.2 (2020): 26.

Pick, J. B., Turetken, O., Deokar, A. V., and Sarkar, A. 2017. "Location Analytics and Decision Support: Reflections on Recent Advancements, a Research Framework, and the Path Ahead," *Decision Support Systems* (99), pp. 1–8. (https://doi.org/10.1016/j.dss.2017.05.016).

Smallman, H. S., and Rieth, C. A. 2017. "ADVICE: Decision Support for Complex Geospatial Decision Making Tasks," in *Virtual, Augmented and Mixed Reality: 9th International Conference, VAMR 2017*, Vancouver, BC, Canada: Springer International Publishing, pp. 453–465. (https://doi.org/10.1007/978-3-319-57987-0_37).

Tigges, C., Hollinsworth, O. J., Geiger, A., and Nanda, N. 2023. *Linear Representations of Sentiment in Large Language Models*, arXiv. (http://arxiv.org/abs/2310.15154).