Data Sheet 1

Ritu Patel, Sarthak Siddhpura, Moin Vinchhi, Vrunda Patel, Vishv Boda
Ahmedabad University, Ahmedabad, Gujarat 380009, India

I. Research Contributions by Paper

- Provided a detailed dataset structure for vehicle trajectory analysis, including fields like CAR_ID, LONGITUDE, LATITUDE, SPEED, HEADING, and TIME, enabling spatial-temporal feature extraction for driving behavior studies.
- Presented driving behavior indicator data (e.g., average speed $\bar{v}$, maximum acceleration $a_{\max}$, standard deviation of acceleration a_s) across multiple samples, offering a statistical basis for identifying normal and abnormal driving patterns.
- Included a principal component analysis (PCA) matrix before and after rotation, demonstrating dimensionality reduction techniques to capture key variance in driving behavior features.
- Supplied Hopkins statistics to assess the clustering tendency of driving data at different speed ranges (0-60 km/h and above 60 km/h), supporting cluster-based anomaly detection approaches.
- Offered emission rates linked to specific power (VSP) intervals, providing an additional dimension (environmental impact) that could correlate with driving behavior anomalies.
- Visualized trajectories and decision tree models (Figures S1, S3), suggesting the use of machine learning techniques to classify driving behaviors based on trajectory data.

II. Key Insights From Paper

The supplementary material provides a rich dataset and analytical framework that align closely with our research direction of identifying abnormal driving behavior using trajectory datasets, leveraging probabilistic and statistical methods. The detailed field descriptions in Table S1 (e.g., LONGITUDE, LATITUDE, SPEED, HEADING, TIME) offer a robust foundation for constructing time series representations of vehicle trajectories on a specific road patch. This is critical for our assumption that drivers exhibit similar behavior in the same spatial context, yet anomalies exist. For instance, the SPEED and HEADING data can be used to compute spatial-temporal features such as velocity vectors or directional changes, which could be clustered to define normal behavior. The driving behavior indicators in Table S3—such as average speed ($\bar{v}$), acceleration (a), and their statistical derivatives (e.g., a_s, $\bar{a}_+$)—provide a quantitative basis for distinguishing normal from abnormal trajectories. By applying clustering techniques (e.g., k-means or k-Shape) to these features, we can establish a baseline of typical driving patterns, enabling a binary classifier to flag trajectories with significant deviations as abnormal, such as those with extreme acceleration ($a_{\max}$) or erratic speed variations (v_s).

A key insight is the use of principal component analysis (PCA) to reduce the dimensionality of driving behavior features, as shown in Table S4. The pre- and post-rotation matrices highlight how variables like speed and acceleration contribute to principal components, suggesting that a few dominant features (e.g., $v_{\max}$, a_s) capture most of the variance in driving data. This is valuable for our project, as trajectory datasets often involve high-dimensional data (e.g., multiple spatial and temporal measurements per trip). PCA could simplify our feature space, enhancing the efficiency of a binary classifier by focusing on the most informative attributes. The Hopkins statistics in Table S5 (0.816 for 0-60 km/h, 0.829 for 60+ km/h) further indicate a strong clustering tendency in the data, supporting our hypothesis that normal driving behaviors can be grouped spatially and temporally. This statistical measure reinforces the feasibility of using unsupervised learning to define normal clusters, followed by a probabilistic approach (e.g., Gaussian Mixture Models) to assign anomaly scores to trajectories deviating from these clusters, aligning with our prob-stat focus.

Additionally, the supplementary material's inclusion of emission rates tied to specific power intervals (Table S6) and projected driver data (Table S7) offers an innovative angle for anomaly detection. While our primary goal is behavioral classification, linking driving patterns to environmental impact (e.g., high CO or NOX emissions) could serve as an auxiliary indicator of abnormal behavior, such as aggressive acceleration or braking. The decision tree visualization (Figure S3) suggests a supervised learning approach to classify behaviors, which could complement our unsupervised clustering by providing a labeled validation step. For our research, we could extract features from trajectories, apply PCA for dimensionality reduction, cluster them to define normalcy, and then use a decision tree or logistic regression to classify anomalies probabilistically. The trajectory diagram (Figure S1) underscores the spatial context of driving, reinforcing the need to segment data by road patches, as per our assumption. This multi-faceted approach—combining statistical feature extraction, clustering, and classification—positions the supplementary material as a practical resource for building a robust anomaly detection system tailored to driving behavior analysis.

III. Unaddressed Issues and Assumptions

- **Unaddressed Issues:**

– The supplementary material lacks details on the methodology for collecting or processing trajectory data, limiting insights into data quality or preprocessing steps relevant to anomaly detection.
– It does not specify how abnormal behaviors are defined or identified, leaving the transition from clustering to binary classification unclear for our purposes.
– The impact of missing or noisy data (e.g., GPS inaccuracies) on the statistical indicators or PCA results is not discussed, which is critical for real-world trajectory analysis.
– Real-time applicability of the proposed features or methods is not addressed, a key consideration for dynamic driving behavior monitoring.

- **Assumptions Made:**
  – Assumes trajectory data is complete and accurate (e.g., consistent sampling rates for TIME, RECEIVE_TIME), which may not hold in practice due to signal loss or delays.
  – Assumes that statistical indicators (e.g., v_s, $a_{\max}$) sufficiently capture abnormal behavior, potentially overlooking contextual factors like road conditions or traffic density.
  – Assumes clustering tendency (via Hopkins statistics) translates directly to meaningful behavioral clusters, without validating cluster interpretability for anomaly detection.
  – Assumes emission rates correlate strongly with driving behavior anomalies, which may not always be true across all vehicle types or conditions.

IV. MOTIVATION BEHIND CHOOSING THIS PAPER

The motivation for selecting this supplementary material lies in its provision of a comprehensive dataset and statistical insights directly applicable to our research goal of identifying abnormal driving behavior using trajectory datasets. The detailed trajectory fields (e.g., LONGITUDE, LATITUDE, SPEED) and derived indicators (e.g., $\bar{v}$, a_s) offer a ready-to-use framework for extracting spatial-temporal features, aligning with our prob-stat approach to model driving patterns on road patches. The PCA analysis and Hopkins statistics provide tools to handle high-dimensional data and validate clustering tendencies, essential steps in defining normal behavior before classifying anomalies. Additionally, the inclusion of emission data introduces a novel perspective that could enrich our anomaly detection by linking behavioral outliers to environmental outcomes. Given our project's focus on developing a binary classifier, this material's emphasis on quantifiable driving features and visual aids (e.g., trajectory diagrams, decision trees) inspires a hybrid methodology—combining unsupervised clustering with supervised classification—making it a valuable resource to advance our research objectives efficiently.

REFERENCES