International Journal of Computer Applications Technology and Research
Volume 14–Issue 02, 113 – 120, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1402.1008

Data Validation and Matching Algorithm for Road Safety Database Integrity in the Philippines

Maywadee Soytong
Scientific Computing Laboratory, Department of Computer Science,The University of the Philippines Diliman,Quezon City, Philippines

Arse John P. Salison
Intelligent Transportation Systems Laboratory, National Center for Transportation Studies,
The University of the Philippines Diliman,
Quezon City, Philippines

Patrick Rollan
Intelligent Transportation Systems Laboratory, National Center for Transportation Studies,
The University of the Philippines Diliman,
Quezon City, Philippines

Adrian Roy L. Valdez
Scientific Computing Laboratory, Department of Computer Science,
The University of the Philippines Diliman,
Quezon City, Philippines

Susan Pancho-Festin
Computer Security Group, Department of Computer Science,
The University of the Philippines Diliman,
Quezon City, Philippines

Harvey Ian Arbas
Intelligent Transportation Systems Laboratory, National Center for Transportation Studies,
The University of the Philippines Diliman,
Quezon City, Philippines

Abstract: Fragmented road safety data across Philippine agencies hinders practical analysis and intervention. This paper explores a data matching algorithm for the Philippine Integrated Traffic Incident Database (PITAD), designed to consolidate data from various sources. The algorithm prioritizes matching incident time/date, and GPS coordinates with secondary criteria such as vehicle types and demographics. By pre-processing data, computing matching scores based on time, location, and additional factors, and clustering matched entries, the algorithm tackles data fragmentation and improves data quality within PITAD. This paves the way for comprehensive analysis of road safety trends, enabling data-driven decisions for targeted interventions and, ultimately, enhanced road safety in the Philippines.

Keywords: Data Validation, Data Matching, Data Integration, Database, Road Safety, Data Storage

1. INTRODUCTION

The Philippines faces a severe road safety crisis. In Metro Manila alone, the Metropolitan Manila Development Authority [1] documented over 71,891 crashes in 2022, with a tragic 424 fatalities. This translates to nearly 200 daily crashes and one death per day [1]. Beyond the human cost, road accidents inflict a staggering economic burden exceeding PHP 3.4 million per fatal incident [2].

Addressing this issue requires a data-driven, multi-sectoral approach, involving the public health, transportation, and education sectors. Despite this need, the fragmented data across multiple agencies hampers analysis and collaboration. This current method of handling data hinders evidence-based policymaking and targeted interventions.

The Philippine Integrated Traffic Incident Database (PITAD) seeks to address this fragmentation. PITAD consolidates data from disparate sources like the Department of Transportation (DOTr), MMDA, Philippine National Police (PNP), and Department of Health (DOH) into a unified system using advanced data matching algorithms to improve data quality and accessibility. Crashes unrecorded by one agency might be captured by another, resulting in a more comprehensive road safety record. This enables transportation authorities to pinpoint accident hotspots, identify recurring patterns, and correlate incidents with factors like weather or infrastructure deficiencies. These data-driven insights pave the way for targeted interventions like improved signage, strategic enforcement, and evidence-based policymaking, paving the way for enhanced road safety management.

2. REVIEW OF RELATED LITERATURE

2.1 Challenges posed by fragmented data systems

Fragmented data systems pose a significant challenge to data analysis and utilization [3]. Scattered information across various databases and applications makes it difficult to access all relevant data for a complete picture. It can lead to incomplete datasets, hindering a holistic understanding of the issues. Inconsistencies in data formatting, storage conventions, and reporting practices across different systems compound the issue, multiplying errors and reducing overall data quality. Furthermore, fragmented data systems impede collaboration

among stakeholders who rely on shared information, restricting the exchange of valuable insights and knowledge [4].

In the Philippines, this challenge is particularly evident in road safety data, where agencies like the MMDA, DOTr, and PNP maintain separate records of crashes, infrastructure, and traffic patterns. These silos create data gaps, inconsistencies, and redundancy, making it difficult to accurately identify high-risk areas, analyze patterns of recurring incidents and contributing factors or implement effective interventions.

Moreover, the lack of a unified approach causes valuable insights to remain trapped within each individual system. Poor data integration leads to missed opportunities for resource allocation and intervention design. To address this issue, centralized and standardized data collection frameworks are essential for a holistic understanding of road safety challenges.

2.2 Available algorithms for data matching

Data matching is identifying records that refer to the same entity across different datasets. It is a crucial step in data cleaning and integration. Various algorithms have been developed to tackle this challenge, and choosing the most suitable approach depends on the data's nature and the desired level of accuracy [5].

2.2.1 Deterministic Matching

This rule-based approach relies on predefined rules that compare specific fields (e.g., name, address) across datasets [5]. Exact matches are identified when all compared fields align perfectly with the stipulated rules. This method is fast, computationally efficient, and well-suited for structured data with high-quality attributes. However, due to strict criteria, they can be overly rigid, failing to account for variations in formatting, abbreviations, or typographical errors, potentially missing valid matches [5].

2.2.2 Probabilistic Matching

In contrast, probabilistic matching algorithms employ statistical techniques to assign probability scores to potential matches based on similarities across various fields [7]. String similarity algorithms like Levenshtein distance measure the degree of similarity between two strings, allowing for variations in spelling or formatting [7]. Probabilistic matching methods are more flexible and can handle inconsistencies in data. However, it requires setting appropriate thresholds for acceptable match scores and may generate false positives due to chance similarities [7].

2.3 Road Safety Databases in the Philippines

Like many nations, the Philippines faces a significant challenge in ensuring effective road safety data management and analysis. Various agencies, including the DOTr, PNP, and MMDA, maintain separate databases for crash reports, traffic incidents, and road infrastructure data [8]. However, this fragmented approach hinders a comprehensive understanding of accident trends and contributing factors.

Previous efforts to integrate road safety data in the Philippines have needed to be expanded in scope and efficacy. The Metro Manila Accident Reporting and Analysis System (MMARAS), the Crime Information Reporting and Analysis System (CIRAS), and the Traffic Accident Recording and Analysis System (TARAS) offer centralized databases for incidents within their respective jurisdictions [9]. However, these systems operate in silos, lacking integration with other relevant data sources and additional information such as weather and road characteristics. Moreover, inconsistencies in data reporting practices, format variations, and incomplete attribute coverage across different agencies further compound the challenges of data consolidation and analysis [9,10]).

The literature review highlights the pressing need for an integrated approach. Innovative data matching algorithms and centralized database architecture could offer a novel solution to overcome the limitations of fragmented data systems and enable comprehensive analysis, inter-agency collaboration, and, ultimately, more effective road safety policies and interventions. Given that PITAD wants to become a catch-all database for all road safety data, the problem of duplicate data becomes unavoidable.

Without a clear protocol for collaboration among agencies that collect and record road crash data, data duplication becomes a problem. For instance, MMDA may have recorded incident data, and ONEISS may also record the same incident through the hospital that has handled the post-crash care. When information is sourced from agencies with potentially inconsistent reporting practices, a data matching algorithm should be developed for PITAD to function and give correct insights properly.

A significant challenge in consolidating road safety data in the Philippines stems from the varying data field structures employed by different agencies. Table 1 shows the current databases that were incorporated into PITAD, as well as the availability of these databases. Databases that the researchers were to access online can have near real-time integration with PITAD. Only those that are integrated with PITAD once the agency responsible sends their data. In addition, the inconsistencies between data fields necessitate data transformation and mapping efforts to ensure compatibility with the centralized PITAD data storage system. Due to this, not all data fields from the source database will be incorporated into PITAD. For instance, fields that do not describe the traffic incident will not be included in PITAD. However, each entry in the PITAD database has the source database indicated and the specific record ID. While PITAD utilizes a standardized 53 field structure, it further incorporates four significant data collection tables: the data collection table, the person table, the vehicle table, and a data match storage table. This structured approach facilitates data organization and retrieval but requires effective data mapping and integration strategies to bridge the gap between the diverse agency data formats and the PITAD structure.

Section 3.3 describes in detail the process of integrating different databases into PITAD, while Section 3.1 describes the data-matching process employed by PITAD.

Targeted interventions could be strategically deployed by correlating these patterns with weather data and specific infrastructure features. These include improved signage, the deployment of enforcement patrols in high-risk areas during adverse weather conditions, or even infrastructure upgrades to address identified safety deficiencies. Ultimately, a unified data management system would empower data-driven decision-making in road safety. This approach has the potential to significantly reduce the number of road accidents and fatalities across the Philippines, leading to a demonstrably safer transportation system.

International Journal of Computer Applications Technology and Research
Volume 14–Issue 02, 113 – 120, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1402.1008

Table 1. Existing Databases incorporated in PITAD

Database	No. of Data Fields	Availability
DRIVER	42	Online database
MMARAS	40	Offline copy
CIRAS	73	Offline copy
ONEISS	44	Offline copy

3. METHODOLOGY

Data matching is done on the PITAD central database, where entities from several source databases (CIRAS, MMARAS, DRIVER) are pre-processed before being added. The pre-processing details are described in Section 3.1.A. The essential fields used for data matching are incident date and time and the incident's GPS coordinates (latitude and longitude). Another set of fields was used as secondary criteria for data matching. Those fields are vehicle types, and the age and gender of the persons involved.

3.1 Design and Development of the Data Matching Algorithm

The data matching algorithm is run once the PITAD database is populated with data from various databases.

3.1.1 Technical description of the algorithm

During the data matching, the PITAD database is matched by itself. However, if the data matching algorithm had been run previously, only the entities entered in the PITAD database after the previous run would have been matched. For each entity that would be matched, entities that are in the neighborhood of that entity are fetched. A data matching score is computed by comparing the timestamp (Unix time), latitude, and longitude. Depending on the source database, additional fields like vehicle types, age, and gender of persons involved were used as additional variables in the computation of the data matching score. The data matching score for the two entities being compared is the weighted mean of the timestamp, latitude, and longitude scores and the additional variable, if there are any. The data matching score is computed by looking at the value's distance to the target value's neighborhood. If the value falls inside the neighborhood of the target value, then the data matching score for that variable is 1. Otherwise, it decreases linearly as distance from the interval bounds increases, with a minimum score of 0. The formula for computing data matching score is disclosed in 3) Section 3.3.2.

3.1.2 Rationale behind the algorithmic choices

This section aims to justify certain major decisions behind algorithmic choices. First is the usage of central databases. Data matching is usually done between 2 databases or with itself, in the case of data deduplication. For multiple databases, there is no way to perform data matching other than pairwise data matching between all possible pairs of databases and then consolidating the results. Another potential problem with data matching with multiple databases is that the user would have to vary the fields for data matching for each pair of databases. One approach to solve this problem is to introduce a unified database that would accept entries from multiple databases to be matched, add a field that keeps track of the source database, and perform data matching with itself (data deduplication) in order to flag the duplicates. The duplicates will then form a cluster, which represents entries from those databases that are matched.

A considerable advantage of this approach is that it could be extensible should another database be incorporated. Since the fields for the central database are already established, it would be easier for government agencies and stakeholders to disclose the required information when negotiating incorporation into the central database. However, one disadvantage of that approach is the need to pre-process the databases individually to fit the central database's fields. Some of the fields can be left empty, but for essential data matching fields like latitude and longitude, a geocoding service was used to obtain those should those data be absent from the source database.

Second, date, time, and coordinates are primary data-matching fields. It is essential to establish that a single entry in the central database corresponds to a traffic incident. Also, since multiple databases have varying fields being considered, selecting the minimal number of fields necessary to identify a traffic incident is essential, as selecting more fields than necessary might lead to a problem with empty fields. Hence, the minimal information needed is time and location. Those fields also benefit from being present in all source databases.

However, several problems have come up with using address data. One problem is that there can be huge inconsistencies with how addresses are recorded between various databases, not only with the format but also with granularity (i.e., One database might only record addresses up to district level, while others might record up to street level, while others might only record the precinct where the incident is reported). This could be addressed by considering the location's neighborhood, not exact addresses. Given that it is hard to mathematically determine the neighborhood of a given address, it is decided to use the latitude and longitude of the location instead. One disadvantage of this approach is that not all databases record the latitude and longitude, and a geocoding service must be used to fill the gap.

Finally, the data matching scoring looks at the distance between the value and the neighborhood of the target value instead of the distance between the two values. One prime consideration for this is that similar to location, there could also be discrepancies between how time is recorded between various databases, and to allow for that margin of error, an appropriate neighborhood for the time variable is set.

3.2 Limitations of the Approach

There are several sources of potential biases in this algorithm. The first is due to the jurisdiction of the databases provided. For instance, MMARAS, which is under MMDA, does not record incidents outside the place of their jurisdiction. Hence, its database heavily leans on data points in Metro Manila. Other databases provided by government agencies do not record or redact personal information, and these limitations must be factored into data matching.

In addition, there are also potential sources of errors in the algorithm:

1. Inaccuracies present in the original databases will be reflected in the data-matching results.
2. As the algorithm relies on a geocoding service to fill up latitude and longitude coordinates that are not present in the original database, there could be induced error when the geocoding service throws a wrong coordinate or does not return any, which means that it would be omitted from the central databases.

3. Verifying whether entries are correctly clustered requires additional context data not present in the central database and must be vetted by authorities.

In this case, a manual verification module was also developed to address that gap.

3.3 Data Processing Workflow

The data processing workflow begins with the extraction of data from various sources. Once extracted, the data undergoes a rigorous matching process to identify and pair corresponding records. Subsequently, filtering techniques are employed to refine the dataset, eliminating irrelevant or low-quality data. Verification steps are implemented to validate the accuracy and consistency of the remaining data and ensure data integrity. Multithreading is utilized to parallelize the processing tasks, significantly improving efficiency and reducing processing time. The compute_datamatch function, a core component of the workflow, leverages a specific algorithm to calculate similarity scores between data pairs. This algorithm, tailored to the specific data characteristics and matching criteria, plays a crucial role in determining the quality of the matched data.

3.3.1 Extracting data from various datasets

PITAD gets its data from CIRAS, MMARAS, and DRIVER, as well as the data entered through the companion app. Various MoAs (memorandum of agreement) were signed with government agencies responsible for keeping these databases to obtain access or soft copies of databases (usually in the form of Excel files). These databases might not share a common identifier, and those that do might have been written in different formats.

A database, called the PITAD database, would be constructed to house entries from these existing databases. First, a set of fields for the PITAD database is identified. After that, a minimal subset of those fields used for data matching is determined. The essential fields that would be used for data matching are incident date and time and GPS coordinates (latitude and longitude). In addition, another set of fields would be considered as secondary criteria for data matching. However, these fields can be empty. The fields essential for data matching cannot be null or empty, while others in the PITAD database that are not essential can be left empty, as the source database might not have the relevant information.

The PITAD database has three tables where data on traffic incidents are stored:

- a DataCollection table where the general details of the incident are stored.
- a Person and Vehicle table where the details of the person(s) and vehicle(s) involved in the incident are stored separately.
- a DatamatchStorageCollection table where the clustering of entities in DataCollection is stored separately

The Person and Vehicle tables are linked to the DataCollection table using a Foreign Key.

After identifying the fields for the PITAD database, the next step is to plan the integration of different existing databases. Some databases are available online (i.e., DRIVER), while others are offline, with Excel files being provided by different agencies. In this case, for those databases that are available online, a script was written to fetch data from those databases through their API and process them to fit the PITAD database. The same strategy would also be used for offline sources. Since each database differs, a customized script would be written for each database to process them and convert them to entities that would fit the PITAD database. However, as the Excel files are not provided regularly, the scripts are run externally, and the resulting files are then manually uploaded into the PITAD database.

In some fields (i.e., weather conditions and light conditions), entries from the source database were mapped into the approximate equivalent accepted by the PITAD database. Some databases do not have complete information for vital fields like longitude and latitude. In those cases, the script includes a function that calls on a geocoding service API to fetch the GPS coordinates given the address or location text. Currently, nominatim is the geocoding service used as it is free. However, in some cases, the service might not be able to get the coordinates of the given address. In most cases, the given address text needs to be more specific, and those entries cannot be entered into the PITAD database.

Each entry in the source database corresponds to a single entry in the PITAD database. Duplicates within the same database or other databases are still entered as separate entries in the PITAD database. An additional field in the PITAD database called source indicates the original database the entity came from.

3.3.2 Data matching process

1) Multithreading

As the PITAD database contains thousands of entries, multithreading will be used to perform data matching to speed up the process.

Let D be the set of entities to be matched.

Split D into several chunks. The number of chunks depends on the number of threads. After that, feed each chunk in the init_data_process function. The init_data_process function handles the data matching of a given chunk. After all threads are finished, save the date and time of the last time a data match was successfully performed on the LastDatamatchPerformedOn table.

2) Init_data_process function

Let d be the data chunk that will undergo data matching. The pseudocode below describes the init data process function.

```
*For each i in d, do the following:
    * Compute the neighborhood parameters for i.
    * Neighborhood parameters for i:
        * Time neighborhood:
          [i.date_time – SEARCH_TIME_OFFSET,
          i.date_time + SEARCH_TIME_OFFSET]
        * Latitude neighborhood:
          [i.latitude – GEO_OFFSET,
          i.latitude + GEO_OFFSET]
        * Longitude neighborhood:
          [i.longitude – GEO_OFFSET,
          i.longitude + GEO_OFFSET]
    * Filter entities in the database
      whose date_time, latitude and longitude are all within
      the time, latitude and longitude neighborhood computed
      previously.
      Let queryset be the set of those entities.
    * Initialize i.parent_id = i.record_id
        * If len(queryset) > 1:
        * Initialize best_match_rate to 0.
          //This variable saves the highest data match rate so far.
```

* For each candidate in queryset:
* If candidate = i:
* Go to next iteration
* datamatch_result = compute_datamatch (i, candidate)
* If datamatch_result.data_match_rate > MATCH_THRESHOLD:
* If candidate is in DatamatchStorageCollection and datamatch_result.data_match_rate > best_match_rate:
* Set i.parent_id = candidate.parent_id
* Set best_match_rate = data_match_rate
* else:
* Go to next iteration
* end for
* Create an entry for i in DatamatchStorageCollection:
* end for

3) Compute_datamatch function

Let i, candidate, be the entities being compared, and x_i and x_{cand} are the values compared.
As interval comparison will be used to compare two values, define K as the acceptable offset. K specifies the acceptable distance between the two values. If the distance between the two values is within K, the data match score will be 1. If the distance between 2 values is greater than K, then the data match score decreases linearly as the distance increases, with the lowest score being 0. Let $K_{TIME}$and K_{GEO}be the K for timestamp and geographic coordinates (i.e., latitude and longitude), respectively.

The data matching score for a single variable is computed as follows:
* Compute bounds of the interval.
Set a = x_i– K and b = x_i+ K
* Compute tol = offset - K
* Compute dist = max(a - x_{cand}, x_{cand}- b)
* a is the minimum values of the interval,
b is the maximum values of the interval,
while dist is distance of value from interval.
* if diff < 0:
* x = 1
* if diff > to l:
* x = 0
* else:
* x = 1 - | dist/tol |
where x is data matching score and offset is the size of the neighborhood
\(i.e.: SEARCH_TIME_OFFSET, GEO_OFFSET)

The algorithm for computing data match score for i; candidate is as follows:

* Compute data matching scores for timestamp, latitude and longitude.
Denote those scores as x_{TIME}, $x_{LAT}$and x_{LNG}
* Check the source database for i and candidate.
* If i.source = candidate.source = DRIVER:
* If all non-redacted vehicle types for i and candidate are equal:
* x_{EXTRA}= 1
* Else:
* x_{EXTRA}= 0
* x = $w_{TIME}x_{TIME}$ + $w_{LAT}x_{LAT}$ + $w_{LNG}x_{LNG}$ + $w_{EXTRA}x_{EXTRA}$
* Elif i.source = candidate.source = CIRAS:
* If all non-redacted age and gender for i and candidate are equal:
* x_{EXTRA}= 1
* Else:
* x_{EXTRA}= 0
* x_{SCORE}= $w_{TIME}x_{TIME}$+ $w_{LAT}x_{LAT}$ + $w_{LNG}x_{LNG}$ + $w_{EXTRA}x_{EXTRA}$
* Else:
* x_{SCORE}= $w_{TIME}x_{TIME}$+ $w_{LAT}x_{LAT}$+ $w_{LNG}x_{LNG}$

Where $x_{SCORE}$is the data matching rate of i and candidate, and is the weighted mean of data matching scores for time, latitude, longitude, and extra variable, with weights w_{TIME}, w_{LAT}, w_{LNG}, and w_{EXTRA}, respectively, and $w_{EXTRA}$is computed as all-or-nothing, and is only computed when comparing entities from specific databases.

4. RESULTS AND DISCUSSION

This section presents the results and discussion of the data matching and clustering process within the PITAD system. We delve into identifying data clusters based on shared parent IDs, followed by a validation process to assess the accuracy of the matching algorithm. The impact of data integration on road safety data integrity is also explored, highlighting the benefits of a more comprehensive and accurate dataset.

4.1 Identification of Data Clusters

After the data matching algorithm is run, as described in the previous section, each entity in the central database will have its assigned parent ID. Entities in the same database with the same parent ID could be due to duplicate entries (if they belong to the same database) or entities from different databases that point to the same traffic incident. Either way, these database entries can be grouped. Hence, a cluster is defined as a set of at least two entries from the central database with the same parent ID. The parent ID can represent a given cluster and be counted as a single object when dealing with traffic incidence counts. A cluster can contain entries that are from several databases. The source field indicates the database from which it came from.

4.2 Validation and Verification of Results

A simple validation was conducted to assess the effectiveness of the data-matching algorithm employed in PITAD. A sample of 40 clusters was randomly selected, and each cluster was examined to determine the accuracy of the data grouping. The validation process involved comparing the clustered data against the criteria presented in Table 2.

International Journal of Computer Applications Technology and Research
Volume 14–Issue 02, 113 – 120, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1402.1008

Table 2. Criteria and Points for Match Accuracy

Match Accuracy	Points	Criteria
Exact Match	5	within +/- 3 hours within +/- 0.001 degrees the same address the same number of vehicles involved the type of vehicles involved the same number, age, and gender of persons involved
Good Match	3	within +/- 12 hours within +/- 0.0054 degrees At least the same city the same number of vehicles involved the type of vehicles involved at least the same number and, gender of persons involved
False Match	0	greater than +/- 12 hours greater than +/- 0.004 degrees not within the same city not the same number of vehicles not the same number of persons involved

The score's accuracy is taken as the ratio of the total points for all clusters and the perfect score (assuming each cluster was a perfect match), as shown in Equation 1.

$$\frac{\Sigma n P_x}{nP}$$

Where:

P_x is the points assigned to the cluster.

n is the total number of clusters.

P is the highest number of points for a cluster (points for exact match).

There are some limitations to this verification method. The information used to cross-check the validity of a cluster's grouping were fields that had no risk of data privacy issues and were in the database (i.e., location text, persons and vehicles involved, etc.) No extra context was used to determine if the entries in the cluster point to the same traffic incident. Furthermore, the method should have addressed potential omissions of entries that should have been included in clusters.

Despite these limitations, the validation results offer valuable insights into the algorithm's accuracy and highlight areas for improvement. Continued refinement of the data matching process, coupled with enhanced contextual verification mechanisms, will bolster the integrity and reliability of PITAD, further empowering data-driven decision-making for improved road safety outcomes.

Table 3. Accuracy Results for Each Model Considered

Method	Number of Clusters	Accuracy
1	40	34.00 %
2	40	62.50 %

Method 1 relied solely on time, date, and location parameters to match data, while Method 2 integrates additional variables such as gender and vehicle type. There is a significant discrepancy in accuracy between Method 1 and Method 2, with the higher accuracy rate observed in Method 2 implying that including gender and vehicle type as supplementary criteria enhanced the matching algorithm's ability to characterize records that may be referring to the same incident.

The validation results demonstrate the importance of incorporating multiple variables beyond time and location data when matching traffic incident records across databases. Including additional parameters like vehicle types and details about persons involved, the accuracy of matching related incident reports improved substantially from 34% to 62.5%.

This highlights that more than relying solely on spatiotemporal data is required for robust data matching, as incidents occurring in proximity or overlapping time frames may not refer to the same event. Layering in descriptive details about vehicles and persons allows the algorithm to distinguish true matches from false positive cases better.

However, the maximum accuracy achieved of 62.5% with the enhanced method still leaves considerable room for improvement. Additional relevant variables could be integrated to increase matching accuracy if that data is available across the different database sources. Factors like crash specifics, road conditions, direction of travel, and more granular location details may help sharpen the algorithm's ability to link related incidents definitively.

Attempting to increase the accuracy of the data matching model will likely be limited by the availability of data from each database, as currently, most fields are either lacking or empty on some incident records. Data sparsity remains an ongoing challenge, as many fields were lacking or empty across incident records from different databases. Encouraging more complete data reporting from the supply side and implementing advanced missing data handling techniques could expand the variable set available for matching. Nonetheless, the current PITAD approach represents a solid step forward by leveraging an expanded set of parameters to significantly boost result accuracy over approaches based on time and location alone.

It will also be valuable to explore developing a confidence scoring system along with the data matching rather than using discrete match/no-match categorizations. This would enable

understanding the degree of certainty in predicted matches, allowing human reviewers to focus on evaluating fewer clear-cut cases. Altogether, the insights gained through this validation study highlighted both the merits of the current approach and multiple promising paths for continued enhancement of PITAD's data integration capabilities.

4.3 Impact on Road Safety Data Integrity

Integrated data matching offers a powerful tool for enhancing road safety data. It tackles issues like inconsistencies and missing information by combining information from crash reports, weather data, and road infrastructure databases. This process leads to a more accurate and complete picture of road safety trends. Additionally, integrated data matching creates a centralized repository, improving accessibility for analysis and fostering collaboration between stakeholders. These improvements translate into real-world benefits. Authorities can pinpoint high-risk areas and crash patterns, allowing for targeted interventions like improved signage or strategic enforcement. Data-driven policymaking becomes a reality, ensuring policies address the most pressing road safety challenges. Finally, tracking performance through integrated data analysis enables continuous improvement of road safety efforts. Integrated data matching strengthens road safety initiatives by providing a clearer picture and empowering data-driven decision-making.

4.4 Recommendations for stakeholders in road safety management

Road safety stakeholders can unlock a powerful tool: collaboration and standardized data. Consistent data formats across systems minimize errors and enable a clearer picture of road safety trends. Shared data agreements between agencies and research institutions unlock comprehensive datasets for analysis. Investing in integrated data platforms allows stakeholders to work together. Joint working groups and public-private partnerships foster knowledge sharing and leverage real-time traffic data. Fueled by standardized data, this collaborative approach leads to targeted interventions, effective resource allocation, and measurable progress in creating safer roads.

5. CONCLUSION

Due to fragmented data across different government agencies, the Philippines has struggled to analyze road safety trends and implement effective interventions. This critical information, vital for saving lives, resides in silos, hindering its usefulness. A novel probabilistic data-matching algorithm has been developed to bridge this gap for the Philippine Integrated Traffic Incident Database (PITAD).

This algorithm tackles the challenge of fragmented data by employing a flexible, probabilistic approach. It prioritizes matching incidents based on time, location (GPS data), and additional details like vehicle types when available. Unlike deterministic methods with rigid criteria, this approach utilizes statistical techniques to assign scores to potential matches, accounting for inconsistencies and potential errors in the data. The algorithm transforms fragmented data into a cohesive and high-quality resource within PITAD by pre-processing information, calculating matching scores, and clustering matched entries. This paves the way for a comprehensive analysis of road safety trends, ultimately enabling data-driven decisions and targeted interventions to enhance road safety across the Philippines.

6. ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to DRIVER, MMARAS, CIRAS and ONEISS for generously sharing their valuable data, which was instrumental in conducting this research. Their contribution significantly enhanced the quality and depth of this study.

7. REFERENCES

[1] Metropolitan Manila Development Authority (MMDA). 2022. “Metro Manila Accident Reporting and Analysis System (MMARAS) Report MMDA Traffic Engineering Center (TEC) - Road Safety Unit.” (https://mmda.gov.images/Home/FOI/MMARAS/MMARAS_Annual_Report_2022.pdf)

[2] De Leon, M.R.M. and Cal, P.C. and Sigua, R.G. 2005. “Estimation of Socio-Economic Cost of Road Accidents in Metro Manila,” Journal of the Eastern Asia Society for Transportation Studies, Volume 6, 3183-3198.

[3] Winkler, W.E. 2006. “Overview of Record Linkage and Current Research Directions,” Census Working Paper, U.S. Census Bureau. (https://www.census.gov/library/working-papers/2006/adrm/rrs2006-02.html)

[4] Elmagarid and Mohammed, A. H. and Gupta, P. 2012 “Data Integration: Fundamentals,” Springer.

[5] Christen, P. 2012. “Data Matching,” Entity Resolution, Record Linkage, Springer.

[6] Mellaoui, A. and Algorithmique, L. 2015. “Deterministic and probabilistic methods for record linkage,” In Proceedings of the 18th International Database Engineering & Applications Symposium. ACM, 262-267.

[7] Winkler, W.E. 1995. “String comparator metrics and linkage rules for record linkage,” Proceedings of the Section on Survey Research Methods. American Statistical Association, Volume 1, 180-185.

[8] Soytong, M. and Valdez, A. and Pancho-Festin, S. 2023. “Designing Data Modeling for Road Safety Integrated Data Storage for the Philippines,” 15th International Conference of Eastern Asia Society for Transportation Studies (EASTS).

[9] Regidor, J.R.F. 2019. “Current state of transportation data and statistics in the Philippines and opportunities for improvement towards usability,” Transportation Research Procedia. Volume 39, 703-710. 10.1016/j.trpro.2019.07.089.

[10] Miraflor, J. 2019. “Traffic Accident Analysis in the Philippines and Assessment of its Potential ITS V2X Safety System Design,” National Center for Transportation Studies. University of the Philippines Diliman

[11] Federal Road Safety Corps (FRSC). 2021. “Road Traffic Crash (RTC) Report: January-September 2021.”

[12] Hogg, K. &Sampaio, S. 2009. “Analysis of data Integration Algorithms,” University of Manchester.

[13] Israa, L. and Hafez, M.S. and Ismali, M.A. 2021. “Data integration using statistical matching techniques: A review,” Statistics Journal of the IAOS. 37. 1-20. 10.3233/SJI-210835.

[14] Mi, T. and Rajasekaran, S. and Aseltine, R. 2012. "Efficient algorithms for fast integration on large data sets from multiple sources," BMC Med Inform Decis Mak 12 59. 10.1186/1472-6947-12-59

[15] Monge, A. 2000. "Matching Algorithms within a Duplicate Detection System," IEEE Data(Base) Engineering Bulletin.

[16] Rahm, E. and Do, H.H. 2000. "Data Cleaning: Problems and Current Approaches," IEEE Data Eng. Bull 23. 3-13.

[17] Sampaio, S. 2010. "Comparing Data Integration Algorithms,"

[18] Shitta-Bey, O. and Olatunbosun, O. and Bankole, F. and Oni, S. and Okunola, S. 2020. "Road Traffic Crash Data Systems in Lagos State, Nigeria – A Situational Analysis."

[19] Vijay, G. 2020. "Data Integration Report," Massachusetts Institute of Technology.

[20] Yan, K. and Zhao, H. and Pang, H. 2017. "A comparison of graph and kernel-based omics data integration algorithms for classifying complex traits," BMC Bioinformatics. 18.10.1186/s12859-017-1982-4.

[21] Department of Public Works and Highways (DPWH). 2013. "Department Order No.114 Series of 2013: Implementation of the DPWH Traffic Accident Recording and Analysis System (TARAS)."

[22] Department of Public Works and Highways (DPWH). 2004. "Department Order No. 40 Series of 2004: Implementation of the DPWH Traffic Accident and Analysis System (TARAS),"

[23] Department of Health (DOH). 2023. "Online National Electronic Injury Surveillance System (ONEISS) January-March 2023 Surveillance Report," Volume 15 Issue 1.

[24] Winkler, W.E. 2006. "Record linkage in uncertain environments," Springer Science & Business Media.