Linking Police and Ambulance Data: Where are the Gaps?

Boyana Nedelcheva

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2023

Abstract

This work looked to match real-world incidents of stabbings in police records with stab-wound related incidents in ambulance records provided by Police Scotland and the Scottish Ambulance Service. We investigated the following research questions: What are suitable and effective matching criteria for linking related incidents across the police and ambulance datasets? Can we automate these criteria? What is the extent of overlap between the records in the two datasets? Can we uncover potential patterns of underreporting?

Two main approaches for classifying records as matches or non-matches were explored: deterministic and seeded iterative SVM. We found that, while automation of matching criteria was possible, the deterministic approach excelled in both outcome and efficiency. Deterministic classification was therefore used to assess the overlap of the two datasets.

Consistent with prior work, it was found that approximately 44% of knife related injuries were not reported to the police. This suggests that police are unaware of a large proportion of stabbings that are otherwise recorded by ambulance services.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Boyana Nedelcheva*)

Linking Police and Ambulance Data: Where are the Gaps?

Boyana Nedelcheva

Master of Science
Artificial Intelligence
School of Informatics
University of Edinburgh
2023

Abstract

This work looked to match real-world incidents of stabbings in police records with stab-wound related incidents in ambulance records provided by Police Scotland and the Scottish Ambulance Service. We investigated the following research questions: What are suitable and effective matching criteria for linking related incidents across the police and ambulance datasets? Can we automate these criteria? What is the extent of overlap between the records in the two datasets? Can we uncover potential patterns of underreporting?

Two main approaches for classifying records as matches or non-matches were explored: deterministic and seeded iterative SVM. We found that, while automation of matching criteria was possible, the deterministic approach excelled in both outcome and efficiency. Deterministic classification was therefore used to assess the overlap of the two datasets.

Consistent with prior work, it was found that approximately 44% of knife related injuries were not reported to the police. This suggests that police are unaware of a large proportion of stabbings that are otherwise recorded by ambulance services.

Acknowledgements

First and foremost, I would like to thank my supervisor, Harry Schone, at Police Scotland for his support throughout this project and the opportunity to work on such an interesting and meaningful problem. I would also like to acknowledge the invaluable contributions of my dear friend Henry Rennolls, whose proofreading and moral support have been a constant source of help and motivation.

Table of Contents

Chapter 1

Introduction

A large number of violence victims who seek medical care often do not report crimes to the police, resulting in discrepancies between police and public health records. These discrepancies can leave police with an incomplete view of community crime, posing a significant challenge to those seeking to understand vulnerability and inform policing policies. Recent years have therefore seen a shift towards viewing violent crime as a public health problem as well as a policing one.

The public health approach to policing is one that emphasises the importance of multi-agency collaboration, data sharing and integration. The aim of this approach is to provide community safety services with supplementary data from ambulance services and emergency departments (ED) in an effort to better understand and reduce violent crime [36]. Studies have shown that improved interagency data sharing can reduce the number of victims of violence [4] [38].

For example, one study examined the effects of sharing anonymised ED data with local public safety agencies. It was found that increased data sharing partnerships led to a decrease in violence-related hospital admissions and an increase in the reporting of minor assaults to the police [15]. Another study found that EDs documented twice the number of assaults as their local police department, suggesting that combined data would provide a more complete picture of community violence [20]. Further work integrated ED and ambulance data with police records relating to violent crimes, indicating that ambulance data can serve a similar purpose as ED data [35].

In addition, one study used ambulance data to identify crime hotspots of community violence-related calls that were otherwise unidentified by the police. It was also reported that only a small number of violence victims who called an ambulance were subsequently transported to the hospital [1]. A cross-sectional study that explored the

characteristics of calls to ambulance services similarly observed that approximately one third of callers were not taken to hospital [11]. Collectively, these studies indicate that a large proportion of incidents attended by ambulance services are not recorded by EDs and the police, and that linking police and public health records can help the police prevent crime.

Linking data from different sources is commonly referred to as record linkage. However, data matching, data linkage, data integration, record matching, entity matching, entity resolution, merging, deduplication and reidentification can all refer to the same thing [12]. The aim of record linkage is to compare records, in one or more datasets, and determine whether the compared pairs of records correspond to the same real-world events (matches) or distinct events (non-matches). When records are compared within the same dataset, this is known as deduplication whilst comparison between different datasets is known as data linkage.

If datasets from two different sources share a common identifier, then linking two datasets can be solved using a simple join operation. However, in most situations, no such identifiers are available, meaning that more advanced linkage solutions are required. In this project, we look to link police data related to stabbings with ambulance data related to stab wounds provided by Police Scotland (PS) and the Scottish Ambulance Service (SAS). By matching records that correspond to the same real-world events, we hope to provide an indication of the extent of under-reporting related to knife crime in Scotland. While our current focus centres on stabbings only, we are driven by the potential application of linkage to a range of other contexts, such as mental health. Our core aims are as follows: (1) determine suitable matching criteria for linking the police and ambulance datasets, (2) understand whether it is possible to automate these criteria, (3) identify the extent overlap between police and ambulance records, and (4) use this to understand potential patterns of underreporting.

The contents of this paper are organised as follows. Chapter 2 introduces the primary steps in record linkage and potential methodologies. Chapter 3 outlines our chosen linkage approaches as well as the rationale behind their selection. Chapter 4 presents our findings and analysis. Finally, chapter 5 offers a brief project summary and concluding remarks.

Chapter 2

Background and Literature

This chapter outlines the main steps involved in record linkage: data preparation, indexing, comparison, classification, and evaluation. We start with a brief overview of the linkage process, followed by a detailed exploration of each step involved.

2.1 The Record Linkage Process

The record linkage process consists of five major steps: data preparation, indexing, comparison, classification, and evaluation. In the data preparation step, two datasets are cleaned and standardised to ensure the consistent and compatible formatting of all attributes. Attributes refer to the individual characteristics or fields present in each record such as names, dates, postcodes, etc. The second step, indexing, involves performing a pairwise comparison of all records in the datasets, considering all possible record pair combinations. In practice, a full pair-wise comparison may result in a significant number of record pairs, which poses computability challenges for large datasets. The indexing step therefore aims to reduce the number of comparisons by only comparing records that are likely to refer to matches. The resulting record pairs are compared in the third step, comparison. Comparisons between each record pair are made based on a set of one or more matching attributes that are common to both records. In the fourth step, classification, each compared record pair is classified into the set of matches, **M**, the set of non-matches, **U**, or the set of possible matches, **P**, depending on the classification model used. Manual review of the possible matches in **P** results in their classification into the set of matches or non-matches. In the final evaluation step, the quality of the linked records is assessed using a number of measures [9].

The subsequent sections delve into each step of the record linkage process in more detail, discussing the methodologies, techniques, and challenges associated with each.

2.2 Data Preparation

Perhaps the most time consuming step of the record linkage process is data preparation and pre-processing. Real-world datasets often contain noisy, inconsistent and incomplete information making data cleaning and standardisation an important first step in the linkage process. A lack of high quality data is considered one of the main obstacles to successful record linkage as the success of linkage relies more on the quality of the input data than on the capabilities of the classification technique [10].

The challenges related to data quality and data quality assessment have been covered in depth in the literature [5] [31]. In the context of record linkage, three dimensions of data quality are most relevant: accuracy, completeness, and consistency. Accuracy refers to how accurate the attribute values are, whether it is known how the data was recorded, and whether the data has been verified for correctness. Completeness refers to the number of missing attribute values, the reasons for the missing data, and whether missing components will impact linkage. Finally, consistency refers to how consistent attribute values are across and within the two datasets used for record linkage. Inaccurate or inconsistent data often appears in the form of typographical errors and unlikely or even impossible values. Such data can result in false matches, where unrelated records are erroneously linked together, or false non-matches, where related records are not matched. Therefore, errors must be corrected and the data being used must be put into standard formats. For example, dates can be formatted as YYYY-MM-DD where Y is the year, M is the month and D is the day.

Handling missing values is also crucial as they can lead to bias in classification outcome or patterns that frequently include missing values [9]. Various methods exist for handling missing values, some of which remove the records or attributes with missing values. This, however, may lead to a large loss of information if the number of missing values is significant. Alternative approaches fill in missing values, either manually or automatically, with constants, the mean, the median, or the mode. These methods are only well suited to numerical data. Some missing values can be inferred from other attributes. For example, the sex of a person can often be inferred from their name. Data imputation and rule-based techniques to find the optimal value with which to fill a missing attribute value are also commonly used in linkage projects [29].

2.3 Indexing

When linking two datasets, **A** and **B**, it is most natural to compare each record in **A** with all records in **B**. The pair-wise combination of a record **a** from **A** with a record **b** from **B** is known as a record pair. The total number of record pair comparisons is therefore equal to $|\mathbf{A}| \times |\mathbf{B}|$, where $|\cdot|$ represents the number of records in a dataset. When the datasets are large, comparing all records becomes an impractical and prohibitive task. For example, two datasets each containing 10^6 records require 10^{12} record comparisons.

Assuming that there are no duplicate records in the datasets, each record in **A** can only match with one record in **B**, and vice versa. This means that the maximum number of true matches is equivalent to the number of records in the smaller of the two datasets, $\min(|\mathbf{A}|, |\mathbf{B}|)$. Therefore, while the maximum number of true matches increases linearly, the computational load increases quadratically with the size of the datasets. As a result, the vast majority of comparisons will be between records that do not truly match [9].

To reduce the computational burden of comparing a large number of records, indexing techniques are often used to remove record pairs that likely correspond to non-matches. If a pair of records is not compared due to indexing, it is implicitly assumed that the records in that pair are non-matches. Several indexing techniques, traditionally referred to as blocking, are discussed below.

Standard blocking (SB) is the simplest blocking approach and has been used in record linkage for several decades [14]. A domain expert selects suitable attributes from the data and uses (parts of) their values to form blocking keys. An example of a blocking key is a postcode. Records with the same unique blocking key are grouped together in a single block. SB has two main advantages: it creates non-overlapping blocks, preventing the duplication of efforts during the linkage process, and it has a linear time complexity. However, SB may result in missing links under some circumstances. For example, if blocking is based on incident dates, two matching records that occur at similar times (11:59pm and 12:02am), but on different days, would not be linked. SB is also sensitive to data noise as even a small difference in blocking key, due to typographical variations for instance, can place matching records into separated blocks. Furthermore, the approach does not impose a limit on block size which raises scalability concerns.

Sorted neighbourhood (SN) is an alternative indexing approach that aims to ad-

dress these concerns [19]. The first step is to combine or concatenate the datasets to be linked and generate sorting keys (similar to blocking keys) for each record. The sorting keys are sorted in ascending order for numerical data and alphabetically for string data, aligning the associated records accordingly. A window of fixed size then moves over the sorted records, comparing only records within the window at any step. The underlying assumption of SN is that records with sorting keys that are close to one another in a sorted order are more likely to match.

SN has two main advantages: it results in linear time complexity for record linkage, and it is robust to noise, allowing a small difference in the sorting keys of records that potentially match. SN can also be applied to include a tolerance in numerical value, making it suitable for numerical data. The main drawbacks of this method is that its performance largely depends on the choice of window size and sorting key, which are often difficult to configure.

These drawbacks are handled by performing multiple passes of the core SN algorithm, changing the sorting keys in each pass in order to improve the quality of the approach. The multi-pass sorted neighbourhood approach is one of the most efficient and widely used indexing techniques for record linkage [32]. A number of other indexing techniques that extend upon the above approaches are discussed in [30].

2.4 Comparison

At the core of the record linkage process lies the detailed comparison of record pairs. These comparisons are used to determine whether two records in a pair are a match or a non-match. However, with real-world data, even after the data has been cleaned and standardised, it is likely that attribute values contain variations or errors. Such variations make the comparison of two values difficult as one cannot rely on exact matching. To solve this problem, a number of methods have been developed that allow for the approximate comparison between attribute values.

When comparing records from two datasets, **A** and **B**, a subset of n common attributes $(a_1, a_2, \ldots, a_3)$ is selected for comparison. For each pair of records $r_{ij} = (r_i, r_j)$, where r_i is a record from **A** and r_j is a record from **A**, an attribute-wise comparison is performed. This involves applying a comparison function, denoted as C_k, to each common attribute of the record pair. The comparison function takes the values of the attributes from both records as inputs and produces a numerical similarity score $c_k^{i,j}$ in the range $[0,1]$. This results in a vector of n values, $c_{ij} = [c_1^{ij}, c_2^{ij}, \ldots, c_n^{ij}]$, called

a comparison vector. The set of all comparison vectors is called the comparison space [18].

Comparison functions can vary in complexity and scope, e.g. simple exact string and numerical comparisons, comparisons that take typographical variations into account, specialised comparisons for date and time values, and even distance-based comparisons based on geographical coordinates (longitudes and latitudes) appear in the literature. The specific choice of comparison function depends on the nature and type of the attribute being compared (string, numerical, categorical, etc.). The following briefly outlines several of the most commonly used comparison functions for string, numerical, date and time, and geographical values.

2.4.1 String Comparisons

The simplest string comparison is the exact comparison (EC). The similarity between two string values, s_1 and s_2, is given by:

$$C_{EC}(s_1, s_2) = \begin{cases} 1, & \text{if } s_1 = s_2 \\ 0, & \text{otherwise,} \end{cases} \tag{2.1}$$

where a score of 1 corresponds to exact similarity and a score of 0 corresponds to exact dissimilarity. Two variations of EC have been proposed. The first variation compares only the beginning or end of the string values. The second encodes strings using an encoding function that replaces similar-sounding strings with similar codes. These codes facilitate the comparison between strings with a phonetic or semantic similarity.

Alternative to exact comparisons are partial comparisons. These compute an approximate similarity score between exact similarity and exact dissimilarity. Some partial comparison functions are based on the edit distance which counts the smallest number of edit operations needed to convert one string to another. For example, the most basic edit distance, the Levenshtein distance [27], assigns a cost of 1 to every single character insertion, deletion, and substitution required to convert one string to another. A number of improved variations have been proposed to reduce its time complexity or to allow for different costs for different edit operations.

Other partial string comparisons involve splitting the input strings into sub-strings of length q characters called q-grams or n-grams. Q-gram-based approaches sequentially slide a window of size q over the input strings and count how many of each q-gram appears in both strings. The numerical similarity can then be computed using

various methods [8]. Extensions based on skip-grams [23] have also been proposed, showing improved matching results compared to q-gram and edit distance-based approaches in the presence of cross-lingual spelling variations.

Additional approaches iteratively find and remove the longest common substring (LCS) of a pair of strings [16]. This process is repeated until either no common substrings remain, or the length of the common substring falls below a given threshold. A substring consists of a consecutive sequence of characters and can refer to a prefix or a suffix. A variation of this approach that extracts the longest common prefix (LCP) is particularly valuable when the shared prefix of a pair of strings carries important and informative information.

2.4.2 Numerical Comparisons

Similarly to string values, numerical values can be compared either exactly or partially to allow for variations and errors. One simple method involves defining a maximum absolute difference, denoted as $d_{\max}$. For this method, the similarity between two numerical values, n_1 and n_2, is computed using the function

$$C(n_1, n_2) = \begin{cases} 1, & \text{if } |n_1 - n_2| \leq d_{\max} \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

This function can be modified to allow for partial comparisons by using a linear extrapolation between exact similarity and exact dissimilarity [9]:

$$C(n_1, n_2) = \begin{cases} 1 - \left(\frac{|n_1 - n_2}{d_{\max}}\right), & \text{if } |n_1 - n_2| \leq d_{\max} \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

Other richer functions, as described in the literature [25], involve the use of distances such as the Euclidean distance, which are then normalised through various transformations. These functions are often used when different attributes carry varying levels of weight.

2.4.3 Date and Time Comparisons

There are two main methods one can use to store date and time information: as a string in some variation of the format YYYY-MM-DD or as a numerical value known as a timestamp. Timestamps represent absolute time values and are typically expressed as the number of seconds or milliseconds that have elapsed since a well-defined reference

point, often called an epoch. Unix time is the most common timestamp with an epoch set to January 1st, 1970, 00:00:00 UTC.

Two methods are therefore available for comparing date and time information: string comparisons and numerical comparisons. String comparisons offer flexibility by accommodating differences in formatting and partial matches. However, they often increase complexity in parsing. On the other hand, numerical methods enable the inclusion of tolerance in the comparison. This proves particularly useful when comparing events that may have occurred within a few hours or days of each other. Numerical methods also allow for more efficient comparisons, making them well suited to larger datasets.

2.4.4 Geographical Comparisons

Geographic information can be represented as either an address or geographical coordinates (longitude and latitude). Addresses are typically stored as strings and compared using approximate string measures to handle typographical variations. In such string based approaches, each address element (house number, street name, town, and postcode) is compared individually. However, a problem with these methods is that they do not consider the spatial relationship between two locations. For example, two nearby locations with different names may be treated as separate entities, even though they are physically close to one another.

To address this problem, one can calculate the geographic distance between two coordinate values and use the resulting numerical value for numerical comparison as discussed in Section 2.4.2. The geographic distance between two points is the length of the shortest path that connects the two points along the surface of a sphere. This distance is typically measured in kilometres or miles. The Haversine formula is one of the most commonly used methods to compute the distance between two coordinates on the Earth's surface.

By using coordinate comparisons, the proximity of two locations can be taken into account. This approach is particularly useful when dealing with tasks such as geospatial analysis, mapping, and location based services, where understanding the true physical distance between two points is essential. If unavailable, coordinate information can often be derived from address information using open source resources.

2.5 Classification

When two datasets share common identifiers, such as NHS numbers, linking two records becomes as simple as performing a standard join operation. In practice, however, such identifiers are often unavailable or missing, necessitating the use of more sophisticated methods for linkage. Three broad categories of linkage method exist: deterministic, probabilistic, and modern approaches. These methods, while generally treated as distinct, share similar implications for subsequent steps in the record linkage process. The aim of each is to classify record pairs based on their true match status using their corresponding comparison vectors.

Deterministic linkage, also known as rule-based linkage, involves the use of a set of pre-defined rules for classifying record pairs based on their agreement over a set of matching attributes. The simplest example of deterministic linkage uses exact matching, where two datasets are joined using a shared unique identifier. More complex approaches implement multiple decision rules, typically starting with those that are least likely to result in a false match. Then, more general rules are applied to identify additional true matches. However, this also makes false matches more likely. For example, consider two real-world events that are likely to match if they occur within a 10 minute interval. Increasing the interval expands the potential matches but also increases the risk of false matches.

When using a set of rules to make decisions, one can determine how uncertain a link is by noting at which step the link is identified. However, when there are multiple matching attributes, there can be a large number of possible agreement patterns. For example, three attributes with binary agreement/disagreement would have $2^3 = 8$ possible patterns. Nine attributes, on the other hand, would have $2^9 = 512$ patterns. When partial agreement is allowed, the number of potential patterns increases even further. Defining and ranking decision rules can therefore be a difficult task, as it often relies on the subjective judgment of the linker. This is the issue that probabilistic linkage methods aim to solve [12].

Probabilistic linkage [14] [28] methods can inform the selection of decision rules by associating each pattern of agreement with the likelihood that two records with the same pattern are a match. Simply, these methods rely on two sets of probabilities: m-probabilities and u-probabilities. M-probabilities refer to the probability that two records agree on every matching attribute given that they are a true match. On the flip side, u-probabilities represent the probability of agreement between two records given

that they are a true non-match. The ratio of these probabilities is called a likelihood ratio and is used to form match weights or scores that indicate the likelihood that a record pair is a match [34]. These scores effectively result in a ranking of all possible patterns of agreement for a set of matching attributes. A decision can then be made regarding the match status of each record pair by comparing their respective patterns against a set of thresholds defined by the linker.

The main issue in probabilistic linkage is thus computing the m and u-probabilities. In the absence of training data, the likelihood ratios, and by necessity the m and u-probabilities, have to be estimated and may therefore deviate from those calculated had the true match status been known. The classic method [14] for estimating these probabilities involves solving a set of quadratic equations, however, this approach is based on the assumption that the matching attributes are conditionally independent. In most cases, this independence relation is likely to be violated. For example, if street name, postcode, and country are the chosen attributes, then records that match on post code are more likely to match on street name and country. Under a weaker assumption than conditional independence, [39] proposed the expectation-maximisation (EM) algorithm to estimate the m and u-probabilities. However, different initialisations are often required to achieve good results. Instead of relaxing the conditional dependence assumption, [40] showed that appropriately incorporating conditional dependence into the original probabilistic model yields comparable or improved matching outcomes. Alternative semi-supervised [24] methods are also proposed in the literature.

Modern approaches aim to improve the quality and scalability of record linkage by leveraging advances in machine learning (ML). Machine learning methods have long been used for pattern classification where the aim is to correctly assign patterns to one of a finite number of classes. In the same vein, record linkage aims to assign record pairs to a set of matches or non-matches based on the agreement patterns between each record pair. Therefore, given a set of agreement patterns, ML methods can be used to predict the class that each pattern belongs to.

The majority of these methods are based on supervised learning, which requires knowledge of the true match and true non-match status of each record pair, i.e. training data. If available, training data can be used to train a classification model to classify record pairs into matches and non-matches. However, in practice, training data is often unavailable, making the manual preparation of training examples necessary. This task is laborious and usually results in training data that is not 100% accurate due to human error, bias, or limitations in the available information that prevent one from discerning

the true match status of a given record pair.

Two popular supervised approaches that have been successfully employed in record pair classification are support vector machines (SVM) [7] [26] and decision tree induction [37]. These approaches often outperform deterministic and probabilistic methods and result in better linkage quality as compared to unsupervised approaches. For example, the authors of [13] implement three classification methods: supervised decision tree induction, unsupervised k-means clustering with three clusters for matches, non-matches and possible matches, and a hybrid approach that combines the first two. The hybrid approach involves two steps. First, a subset of comparison vectors are clustered into matches, non-matches, and possible matches. Then, the clusters containing matches and non-matches are used to train a decision tree classifier. The supervised and hybrid approaches were shown to produce better results than the unsupervised clustering approach on both synthetic and real data.

Various unsupervised clustering techniques have been used for automatic record pair classification. One such technique uses the k-means clustering algorithm to separate comparison vectors into matches and non-matches [18]. A region in between the two cluster centroids can be identified which contains record pairs that are difficult to classify. These pairs can then be classified via manual review. This approach was shown to have high linkage quality, while reducing the number of pairs that have to be reviewed. Other approaches have used clustering to improve indexing and blocking by forming blocks based on the records placed in the same cluster [2]. One study [3] uses canopy clustering, an approach that groups records into overlapping clusters using a distance measure. Records within each canopy are then linked, reducing the number of comparisons required.

Active learning approaches aim to overcome the need for training data by selecting the most informative record pairs for manual review. In [33], a small labelled dataset is initially used to train a classifier. After classification, the most difficult to classify record pairs are given to a user to label. These pairs are added to the training data and the classifier is retrained. This process is repeated until all record pairs have been classified. It was shown that manually classifying less than 100 record pairs can achieve better quality results than a supervised method with 7000 randomly selected training examples.

An alternative semi-supervised method for record linkage is explored in [22]. The method initially trains a classification model on a small set of training examples known as seeds. This initial model is used to classify unseen record pairs. A different classi-

fication model is then iteratively trained on a small percentage of the most confidently classified record pairs, repeating the process until all unseen pairs are classified or a set number of iterations is reached. The authors use an ensemble technique known as boosting with Random Forest and Multilayer Perceptrons as the base classifiers to maximise performance on the classification of unseen data.

While semi-supervised methods reduce the number of manually labelled training examples needed to train a classifier, they still require some level of human input. As a solution, [7] propose an unsupervised approach based on automatic self-learning. Their work uses an approach known as nearest based to automatically select seeds. The nearest based approach sorts all similarity vectors by their distance from an exact similarity vector $[1,1,\dots]$ and an exact dissimilarity vector $[0,0,\dots]$. The similarity vectors closest to exact similarity or dissimilarity are then chosen as match and non-match seeds, respectively. Three sizes (1%, 5%, and 10% of the entire dataset) of non-match seeds were evaluated. The number of match seeds were selected based on an estimated ratio of matches to non-matches. The seeds were then used to iteratively train a classifier using a similar approach to that in [22]. It was observed that while the approach outperformed other unsupervised techniques, it struggled to perform well on datasets containing very few true matches.

More recent work builds upon [21] by combining ensemble learning with automatic self-learning and unsupervised field weighting. In this work, an ensemble is created using various similarity measure schemes initially selected using cosine similarity. These similarity measures are used to generate similarity vectors. An unsupervised field weighting method is then used in combination with the nearest based approach to improve seed selection. The final ensemble is selected based on seed diversity. It was shown that this method improves the quality of the selected seeds and thereby results in better classification. However, the proposed approach cannot handle missing data well and requires a significant number of record pair comparisons in order to generate similarity vectors.

2.6 Evaluation

Record linkage involves balancing the need to maximise linkage accuracy and analysis validity with the constraints of limited human resources, computing resources, and data quality. When making decisions about how to classify a given record pair, two main aspects are considered: the likelihood that the method results in false matches

(where two records corresponding to different events are matched) and the likelihood of missed matches (where two records corresponding to the same event were not matched). However, the balance between the two types of linkage error generally depends on the requirements of the analysis.

If we assume that the true match status of a record pair can reasonably be determined from the available information, then each compared and classified record pair can be assigned to one of the four following categories:

- True positives (TP): record pairs that have been classified as matches and are true matches.
- False positives (FP): record pairs that have been classified as matches, but are not true matches.
- True negatives (TN): record pairs that have been classified as non-matches and are true non-matches.
- False negatives (FN): record pairs that have been classified as non-matches, but are in fact true matches.

To be able to evaluate linkage outcomes, these categories must be summarised into a score. The most commonly used measure to assess classification performance, is accuracy. However, in most record linkage problems, the number of true negatives is often much larger than the sum of the true positives, false positives, and false negatives because of the way records are compared. Due to this imbalance, the large number of true negatives dominates accuracy and produces results that are too optimistic, making accuracy an unsuitable measure for evaluating the quality of record pair classification.

Instead, linkage quality is typically a trade-off between two key metrics:

$$\text{Precision} = \frac{TP}{TP+FP}, \qquad \text{Recall} = \frac{TP}{TP+FN}.$$

Precision computes the proportion of all classified matches that have been correctly classified as true matches. On the other hand, recall computes the proportion of true matches that have been correctly classified. Because precision and recall do not include the number of true negatives, they do not suffer from the imbalance problem [9].

Depending on the situation, it might be more important to prioritise results with higher precision than recall, and vice versa. For example, if matching certain suspect individuals with a large database of people, it can be important to identify all possible

matches at the expense of investigating false matches. This requires high recall. On the flip side, high precision is a priority in situations where linkage is used to identify individuals that need to be contacted about a sensitive health issue.

However, neither precision nor recall alone can completely capture the quality of linkage. For example, one can obtain perfect recall if all record pairs are classified as matches. This would result in a low precision because of the large number of false positives. It is therefore common to combine precision and recall into the f-score which calculates the harmonic mean between precision and recall:

$$\text{F-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{2.4}$$

The f-score strikes a compromise between precision and recall as it only returns a high value if both precision and recall are high.

Chapter 3

Method

In this chapter, we provide the details of the work undertaken in this project. The first section provides a description of the data provided by Police Scotland and the Scottish Ambulance Service. Subsequent sections outline the methods used in each step of the record linkage process, including indexing and comparison, classification, and evaluation techniques.

3.1 The Data

Two datasets relating to knife crime and knife wound-related calls were provided for linkage by Police Scotland (PS) and the Scottish Ambulance Service (SAS), respectively. Both data sets span a two year period from 1st January 2021 to 31st December 2022.

Police Scotland Dataset A subset of data containing information on incidents that featured "knife", "blade" or "stab" in the incident description was obtained from Police Scotland. The information is recorded in a service centre where calls are received and incidents are created by the call handler. Officers may also create incidents themselves.

The dataset includes the categories shown in Table 3.1. When a call is made, date and time stamps are automatically recorded, while all other information is either selected or manually inputted. Date and time are therefore the most consistent and complete fields and give a reliable indication of when an incident was reported. Data on postcode and caller number is less consistent, with approximately 0.5% and 8.6% of values missing, respectively. Postcode information is generated by placing a digital pin on a map using details provided by the caller and is generally accurate to the street level.

Variable	**Description**
ISR Number	Automatically generated case ID
Date	Date of incident
Time	Time the call started
Final Service Code	Type of incident[1]
Description	Description of incident
Postcode	Full postcode
Caller Number	Partial caller number[2]

Table 3.1: Table containing some of the variables provided in the PS dataset. Additional variables provided for analysis have not been included. [1]There are 55 unique service codes such as assault, robbery, and domestic incident. [2]Partial mobile numbers consist of the first four digits after the country code (07 or +44). Landlines consist of the first four digits after the area code.

An additional dataset containing more details is also provided, specifically for incidents linked to more than one crime. This dataset is geographically partial, covering only Highlands and Islands, the North East, Tayside, Fife, and Forth Valley. The records in this dataset are linked to the first dataset through their respective ISR numbers.

Scottish Ambulance Service Dataset The SAS dataset contains 2070 records that are selected based on the wound type "penetrating trauma". This data is routinely collected by call handlers from SAS ambulance control centres. The categories shown in Table 3.2 are included in the dataset.

Like in the police dataset, the date and time stamps are recorded automatically and are thus consistent and complete. All other variables are manually inputted or selected from a drop down menu. In this dataset, the postcode sector and caller number fields contain 0.8% and 28.9% missing values, respectively.

The two datasets share five common fields: date, time, description, postcode (sector), and caller number. However, for the ambulance dataset, incident location is only provided at the postcode sector level, which consists of the first part of the postcode before the space and the first digit following the space. For example, the postcode unit ML7 2SX is part of the postcode sector ML7 2.

The postcode information in both datasets covers similar geographical areas and contains few missing values. Of the 886 unique postcode units in the PS dataset and

Variable	Description
Date	Date of the incident
Time	Time the call started
Description	Description of the incident
Postcode Sector	Postcode sector
Wound Issue	Type of wound[1]
Caller Number	Partial caller number[2]

Table 3.2: Table containing the variables provided in the SAS dataset. [1]There are 10 unique wound issues including swelling, abrasion, laceration, penetrating, foreign body, haemorrhage, degloving, bruising, amputation, and haematoma. [2]Mobile numbers consist of the first four digits after the country code (07 or +44) and landlines consist of the first four digits after the area code.

580 unique postcode sectors in the SAS dataset, 570 unique instances match on postcode sector. This makes incident location a good field for linkage. Other suitable linkage fields include date and time as both fields cover the same two year period and contain no missing data in both datasets.

Caller number is another potential linkage field. However, initial record pair comparisons revealed that there were no instances where caller number matched for matching incidents. In other words, individuals never contacted both the police and ambulance services regarding the same incident. When uncertain about which agency to call, individuals may contact one and assume that that agency will coordinate with others as needed. Caller number is therefore not included as a linkage variable. Furthermore, while both datasets contain incident descriptions, the descriptions are recorded as unstructured text, making them inconsistent and thus unsuitable for linkage. The set of fields used for linkage are therefore date, time, and postcode (sector).

3.2 Data Preparation and Pre-processing

3.2.1 Irrelevant Data

A core issue when analysing stabbings and stab wound-related calls is the presence of information for incidents unrelated to stabbings in the datasets, like accidents or self-harm. These incidents, included as a result of the data collection process, could

introduce bias in later analyses and lead to conclusions that do not accurately represent the patterns of knife crime. To address this problem, manual labelling of the two data sets was performed to categorise incidents as related or unrelated to stabbings. The labels are as follows: (1) unlikely, (2) possible, (3) probable, (4) definite, (5) other weapon types, and (6) self-inflicted.

Definite incidents include only those which specify that a person has been stabbed with a knife. For the police data, definite incidents were verified against the additional dataset containing further details if available. Probable incidents include any stabbings that do not specify a weapon type, e.g. "male stabbed", where it is assumed a knife has been used. Possible incidents generally include potential stabbings and claims or statements that one has been stabbed. Unlikely incidents contain incidents in which a knife or blade is mentioned, but a stabbing is not explicitly stated e.g. "threat to stab" or "female with knife". The other weapon types category consists of stabbings that occurred with anything other than a knife or blade (machetes, swords, and hatchets are considered blades). The self-inflicted category includes incidents in which a person has accidentally or purposefully harmed themselves. If a person has harmed themselves using something other than a bladed instrument, then the incident is included in the self-inflicted category.

While the proposed solution attempts to solve the problem of unrelated incidents, it has several limitations. First, manual labelling is a subjective process that can lead to inconsistent categorisations. Even labels assigned by a perfect annotator can be systematically biased due to data quality. Second, human error is likely to introduce misclassifications, especially in large datasets where the time investment is significant and the task laborious.

To mitigate these issues, clear guidelines about the labelling process were established, taking into consideration ambiguous cases. These guidelines were informed by collaboration with the respective agencies. Furthermore, random samples were selected for review to help verify the reliability of the labels. Various automatic approaches for labelling were also considered as alternatives. However, due to the limited time frame of this project, they were not used here. Further work could explore and implement such alternatives.

In the context of linkage, matching across two categories may result in relevant matches or patterns being missed or overlooked. For example, an incident described as "male seen with knife" could reasonably match to another described as "male stabbed". However, the first would be labelled as 1 (unlikely), while the second as 3 (probable).

To ensure that potential matches are not missed, linkage is performed across the entirety of both datasets. Then, only incidents with possible, probable and definite labels are kept for further analyses.

3.2.2 Standardisation and Cleaning

Prior to linkage, the datasets were reformatted and recoded to ensure consistency. This process aligned the data types for all linkage fields in both datasets. Additional data processing tasks are detailed in the following.

Missing values were removed as the subsequent comparison step involves only two fields, namely the combination of date and time, and postcode (sector). The former attribute is complete and the latter contains a very small proportion of missing values, ensuring that information loss is minimised. Setting missing values to 0 in the comparison step would ultimately yield the same result due to the limited number of comparison fields.

False/hoax calls were removed as they provide no useful information for linkage and subsequent analyses. The final service code “false call” and the class codes ”false/hoax calls to emergency services” and ”falsely accusing named person of crime” were used to identify false calls. Class codes are provided in the dataset containing additional details and represent the incident type. In the SAS dataset, relevant information could be located in the incident description. Specifically, hoax calls are described as ”patient not found”. No such calls were found in the SAS dataset.

Duplicates were removed from both datasets to reduce the possibility of false matches. The PS dataset records duplicates in the “final service code” field using the key phrase “duplicate incident” (n=2297). These duplicates are instances in which more than one person has called to report the same incident. If this is found to be the case, the call handler flags the incident as a duplicate. Usual procedures require that the duplicates are linked back to the live incident, but this does not always happen and the data input is inconsistent. For this reason, we chose to simply remove incidents with the duplicate flag from the PS dataset without linking them back to their original counterparts.

On the other hand, the SAS dataset contains two types of duplicates. The first type contains the keywords ”dup” or ”duplicate” in the incident description and occurs when a person re-dials 999 while waiting for an ambulance, resulting in repeated calls (n=27). The SAS does not link the call IDs, making it impossible to identify a dupli-

cate's counterpart directly. It is assumed, however, that the counterparts are provided in the dataset and that their duplicates could therefore be safely removed without loss of information.

The second type of duplication involves the presence of multiple identical records (n=673). These records share the same attribute values for all fields except "wound issue". Each duplicated incident is repeated a different number of times, and the reason for this duplication is unclear. One possible explanation is that wound issues were recorded separately to other incident information along with an incident ID. If some incidents have multiple entries for the wound issue, then merging the wound issues with the original incidents could lead to such duplications. While removing these duplicates may result in the loss of some data related to wound issues, it is important to note that the wound issue attribute is not used for matching. Therefore, it can be safely disregarded.

Date and time values were transformed into Unix timestamps for numerical comparison in subsequent steps. This approach accommodates incidents that happen at similar times but on different days. For example, an incident that occurs at 11:58pm on Friday in one dataset might appear in the other at 12:02am on Saturday.

3.3 Indexing and Comparison

In this step, records are compared based on timestamp and postcode information. The indexing and comparison methods used in each iteration of linkage are detailed below.

Sorted neighbourhood indexing was used to generate record pairs. All records occurring within one day of each other were paired, as informed by previous findings that police and ambulance records are unlikely to match when more than 24 hours apart [35]. Initial tests confirmed that larger time windows increased the number of false matches considerably. This approach reduced the number of record comparisons from over 29 million to just under 120 thousand.

Timestamps were compared using two of the approaches outlined in Section 2.4.2. The first involves a simple binary match/non-match within a specified time window (see Equation 2.2). If the difference between two incident timestamps is less than the time window, then the incidents match. Window sizes of 3 to 24 hours were tested on either side of an incident. This approach was chosen due to its simplicity and ease of interpretation. It also closely aligns with a previous study that linked police and ambulance data based on time and location [35].

The second approach partially matches timestamps using a linear extrapolation between exact similarity and exact dissimilarity (see Equation 2.3). A predefined window of 24 hours was used to denote exact dissimilarity. Specifically, identical timestamps were given a score of 1, timestamps more than 24 hours apart were given a score of 0, and all other timestamp comparisons received scores that decay linearly from 1 to 0. This approach reflects our belief that incidents become less likely to be related as the time interval between them increases. Partial similarity scores are also well suited to machine-learning classification approaches, which we use in our classification step.

Postcode unit and sector information was provided in the police and ambulance datasets, respectively. Postcode sectors represent a higher level of aggregation, while postcode units represent smaller, more granular areas. The difference in granularity makes their comparison an interesting yet challenging problem. We discuss the two approaches used in this work for comparing postcode information.

The first approach used is the exact comparison (EC) of two strings, the simplest and most common approach to comparing postcode information. By truncating the postcode unit information in the PS dataset to the postcode sector level, a direct comparison can be made with the corresponding data in the SAS dataset. The postcode unit and sector attribute values are automatically generated by both agencies and recorded in standardised formats, making these attributes ideal for EC. While this approach is efficient, we recognise that the information in each dataset may differ slightly for the same incident, for example, due to an incident happening close to a postcode sector boundary.

Postcode sectors provide a moderate level of granularity, but they may still cover relatively large areas especially in rural or less densely populated regions. As a result, two reports that are close to one another but fall on opposite sides of a sector boundary might not be linked. Likewise, two distant incidents within the same postcode sector might be falsely linked. While we cannot address the latter point due to lack of more granular location information, we attempt to account for cross-boundary cases using a geographical approach. To mitigate boundary related issues, we propose a method that utilises the shapefiles of postcode sectors and units. Shapefiles contain information related to the spatial features of postcode units and sectors, such as their shapes, boundaries, and centroid coordinates. This information can be used to define buffer regions around each postcode unit. The percentage overlap between the buffered units and postcode sectors can then serve as a measure for assessing the similarity between a given unit and any sectors it touches.

For instance, units located near the centre of a sector will receive a similarity score of 1 when 100% of the unit area and its buffer lie within that sector's boundaries. However, the buffers of units near sector boundaries may overlap with a neighbouring sector. In such cases, if 20% of the buffered unit overlaps with a neighbouring sector, its similarity to that sector would be 0.2. The remaining 80% would lie within the unit's own sector, resulting in a similarity score of 0.8. This approach may account for boundary cases while minimizing the weight of matches to neighbouring sectors, except when the unit is in close proximity to the sector boundary.

3.4 Classification

In this section, we outline the deterministic and machine learning approaches chosen to classify record pairs into matches and non-matches and clarify the rationale behind our selection.

3.4.1 Deterministic Classification

The first classification method we implement is deterministic. This was selected due to its simplicity, reproducibility, and effectiveness in prior research, particularly in linking police and public health sector data [17]. Another reason for this selection is our use of only two linkage fields. This makes the number of possible agreement patterns minimal and allows us to capture true matches while minimising false matches.

Our implementation is inspired by the approach outlined in [35] as it tackles a closely related matching problem using time and location linkage fields. The matching procedure is simple: if two records match on timestamp and postcode information, then the pair is considered a match. Otherwise, the pair is considered a non-match.

Whether two attribute values match is determined in the comparison step. Timestamps were compared using a binary match/non-match within time windows of 3, 4, 5, 6, 7, and 8 hours either side of an incident. Postcode information was compared using exact comparison in the first matching iteration and postcode unit-sector overlap comparison in the second (see Section 3.3). Further details as to matching iterations can be found in Section 4.1. In addition, each record in the PS dataset was matched with one record from the SAS dataset. The records with the highest comparison scores were retained if a duplicate was found.

While deterministic methods have demonstrated efficacy in linkage projects such

as ours, they also necessitate careful manual selection of classification criteria. We therefore explore an alternative approach aimed at automating this process in order to streamline classification.

3.4.2 Iterative Support Vector Machine Classification

To automate the selection of matching criteria, we follow the unsupervised learning approach outlined in [7]. Our motivation for choosing this approach was twofold: first, it was found to outperform traditional unsupervised clustering techniques, and second, its implementation is straightforward and interesting.

The approach used consists of two main steps for record pair classification. In the first step, training examples are automatically selected based on the detailed comparison between attribute values. The training examples are then used in the second step to iteratively train a support vector machine (SVM). There are two main assumptions underlying this approach. First, if two records relate to the same event, then their comparison vectors are expected to have very high or exact similarity in all linkage fields. Second, if two records relate to different events, then their comparison vectors are expected to have very low similarity in all linkage fields. By selecting such comparison vectors as seeds for training data, one can train a binary classifier to classify all vectors into matches or non-matches.

Two different approaches can be used to select training examples: distance thresholds and nearest based [6]. We chose to use the nearest based approach because it was shown to produce better matching outcomes. In this approach, comparison vectors are sorted according to their distances from vectors representing exact similarity, $[1,1,\dots]$, and exact dissimilarity, $[0,0,\dots]$. These distances were computed using the Euclidean distance between vectors. The vectors respectively nearest to exact similarity or dissimilarity were selected as training examples, forming two distinct training sets of matches and non-matches.

The number of examples selected into the non-match set was 10% of all comparison vectors computed during the comparison step. The number of examples selected into the match set was calculated according to the ratio of true matches to true non-matches, which is estimated as:

$$r = \frac{\min(|\mathbf{A}|,|\mathbf{B}|)}{|\mathbf{C}| - \min(|\mathbf{A}|,|\mathbf{B}|)}, \tag{3.1}$$

where $\mathbf{A}$ and $\mathbf{B}$ denote the two datasets to be linked, $\mathbf{C}$ denotes the set of all com-

parison vectors, and $|\cdot|$ denotes the number of elements in a set. Because the number of non-matches is typically much larger than the number of matches, selecting balanced training sets makes it more likely that comparison vectors selected into the match set are not true matches. Therefore, it is better to select more examples into the non-match training set than into the match set. Once the seed training examples are selected, they are used to iteratively train an SVM.

In the classification algorithm, an initial SVM is used to classify all unseen comparison vectors. A small percentage of the most confidently classified vectors is then iteratively added to the training sets of subsequent SVMs. This process is repeated until all vectors have been classified. Two input parameters must be specified: ip and tp. The first parameter determines the percentage of unclassified comparison vectors that are added into the training sets. The second determines the total number of comparison vectors that are added into the training sets. These parameters were both set to 25 as initial tests revealed that they had little influence on linkage results.

3.5 Evaluation

Given that no ground truth data was available, we conducted a manual review of the matching outcomes using all available informational cues. In particular, incident descriptions allowed us to determine whether an incident was likely to be a match or a non-match. When the match status of a record pair was unclear, that pair was forwarded to PS for further review. During our evaluation, we treated likely matches as true matches and likely non-matches as true non-matches.

Due to the large number of record pairs and the project's time constraints, it was only feasible to review small subsets of matches and non-matches. By randomly sampling the set of matches and non-matches and counting the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN) within these sets, one can compute estimates of the precision and recall.

However, recall that the number of true non-matches (TN) is often significantly larger than the number of true matches (TP). Therefore, random sampling from the non-match set is unlikely to capture the relatively small proportion of missed matches (FN), should they exist, making it difficult to estimate the recall accurately. For this reason, we do not estimate precision and recall directly. Instead, we adopt an iterative improvement approach where we systematically test different matching criteria to find a balance between precision and recall, refining our results over each matching

iteration.

Deterministic classification was evaluated as follows:

1. Apply strict matching criteria that is very likely to return only true matches.
2. Count the number of true matches (TP) and false matches (FP) in the set of matches.
3. Relax the matching criteria. For example, increase the time window from 3 hours to 4 hours.
4. Count the number of additional true matches (TP) and false matches (FP) in the set of matches.
5. Repeat steps 3 to 4 until the number of additional false matches is significantly larger than the number of additional true matches.

Several iterations of deterministic classification were performed using different methods for computing the agreement between two attribute values. The methods tested in each iteration were as follows:

- **Iteration 1**: Exact postcode sector comparison and binary match/non-match between timestamps for varying time windows.
- **Iteration 2**: Postcode unit-sector overlap comparison (Section 3.3) and binary match/non-match between timestamps for a fixed time window, as informed by iteration 1.

The focus of iteration 1 was to determine an effective time window within which to match incidents. Time windows of 3, 4, 5, 6, 7, and 8 hours either side of an incident were tested. Longer time periods of up to 24 hours either side of an incident were also tested, but these did not yield improved results and led to an increased probability of matching unrelated events. We therefore exclude these from the results. On the other hand, the aim of iteration 2 was to determine whether it was possible to capture matching incidents that occurred close to a postcode boundary, but on opposite sides. For this reason, a fixed time window, as informed by the previous iteration, was used along with a postcode unit-sector overlap comparison.

Using this approach to evaluation, suitable matching criteria could be selected by balancing the number of additional true positives against the number of additional false

positives. The final matching criteria and the number of matches and non-matches serve as a benchmark against which the performance of the machine learning approach is assessed.

Iterative SVM Classification was evaluated as follows:

1. Train the model using different combinations of parameters and comparison functions, noting the total number of resulting matches and non-matches.

2. Select all combinations that result in a total number of matches that is similar to that achieved by the deterministic approach.

3. Manually review the set of matches for all selected combinations.

Several matching iterations were performed using different methods for generating the comparison vectors:

- **Iteration 1**: A linear extrapolation between exact similarity (a score of 1) and exact dissimilarity (a score of 0) over a 24 hours period was used for timestamp comparisons. Postcode sectors were compared exactly.

- **Iteration 2**: A linear extrapolation between exact similarity (a score of 1) and exact dissimilarity (a score of 0) over a 24 hours period was used for timestamp comparisons. Postcode sectors were compared using the unit-sector overlap approach.

Note that the approach used to compare timestamps is the same across all iterations, whereas the postcode approach is varied. Both iterations 1 and 2 were initially tested using 6 SVM parameter variations: three kernel methods (linear, polynomial and RBF), and two values for the cost parameter (1, 10). The number of matches resulting from each variation in each iteration were used to determine the most suitable parameter settings. The parameters that produced the most similar results to the deterministic approach were selected for more detailed manual review. This is because the results of the deterministic approach provide us with an approximate number of matches to aim for.

Chapter 4

Results and Discussion

In this chapter, we present our results. We start by discussing the deterministic and machine learning approaches to classification and subsequently present an analysis of the resulting data overlap.

4.1 Deterministic Classification

Deterministic classification was used to classify record pairs into match and non-match sets based on their agreement over a set of attributes, namely timestamp and postcode (sector). Several matching iterations were performed as outlined in Section 3.5. The first iteration was used to determine an effective time window within which to match incidents.

Figure 4.1 shows the the number of additional likely false positives and true positives that were counted for varying time windows. Notably, the number of additional false positives consistently increases as the time window expands. This increase is gradual from 3 to 6 hours and becomes more pronounced between 6 and 8 hours.

In contrast, the number of additional true positives diminishes as the time window increases, increasing by one at 4 and 5 hours and then plateauing. As anticipated, this indicates that incidents are less likely to relate to the same event as the temporal distance between them increases. Furthermore, a time difference of 4 to 5 hours between incidents appears to strike a good balance between capturing additional true positives, while minimising false positives. This is in line with previous work.

The second matching iteration was used to capture incidents that occurred close to a postcode sector boundary, but on opposite sides. A fixed time window of 4 hours was therefore used along with a postcode unit-sector overlap comparison. This approach

introduced 12 additional matches that were not captured by the exact postcode sector comparison in a 4 hour time window. In some cases, incidents from neighbouring postcode sectors occurred within a short time frame of each other. This could lead to the assumption that they relate to the same incident, given their spatial and temporal proximity. Indeed, 7 of the 12 incidents shared matching descriptions and were considered true positives.

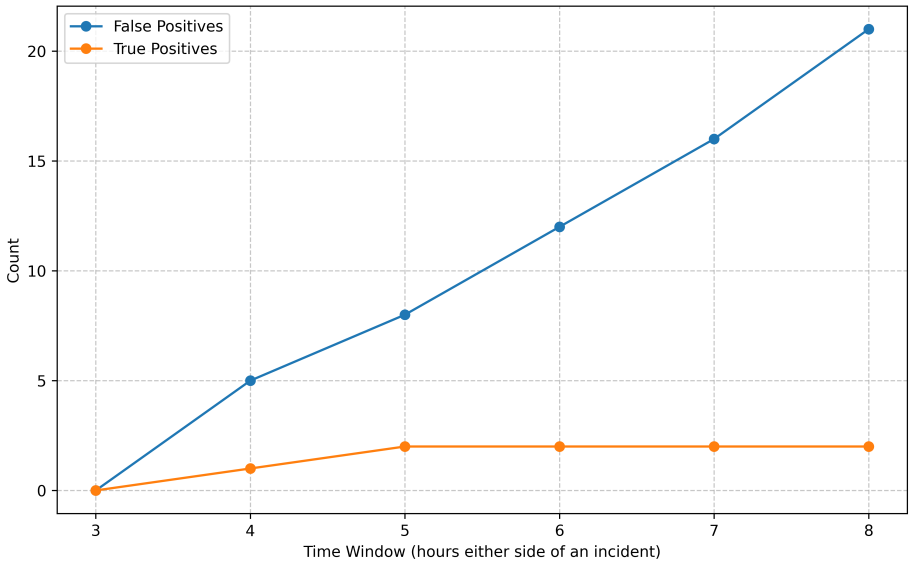

Figure 4.1: The number of additional false positives and true positives with varying time window.

However, challenges arise due to the way boundary cases are accounted for. In our approach, the agreement between postcode information is computed as the percentage overlap between a postcode unit with a buffer zone and a postcode sector. Because of this, an incident that occurred in a postcode unit could potentially be matched with any incident inside the sector it touches, regardless of its specific geographical location within that sector. In other words, an incident that occurred at a sector boundary could be matched with another that occurred in a distant location in the neighbouring sector. For this reason, 5 additional false positive matches are also included within the 4 hour time window.

Interestingly, these false positive matches seemed to be associated with events where the time difference exceeded one hour. This is likely because postcode sectors cover large geographical areas. Therefore, as the time gap between two events

increases, the more likely it is that incidents occurred in different locations. Stricter matching criteria was subsequently applied to exclude events that do not match exactly on postcode sector and occur more than one hour apart.

Drawing from previous iterations, our final iteration classified incidents using a binary match/non-match within a 4 hour time window and the postcode unit-sector overlap comparison. An additional criterion was applied to exclude incidents that occurred in different postcode sectors and more than one hour apart. In line with our expectations, these results suggest that incidents are likely to be related if they occur in close spatial and temporal proximity. Increasing the time window beyond 4 to 5 hours does not yield better results, but it does increase the probability of incorrectly matching two unrelated incidents.

4.2 Iterative SVM Classification

Here we assess whether it was possible to automatically generate matching criteria using an unsupervised learning classification approach. This approach used an iterative SVM to classify record pairs into the set of matches or non-matches. The SVM was trained on seed training examples that were automatically selected using the nearest based approach. The training examples consist of comparison vectors or features generated in the comparison step. Several classification iterations were performed as outlined in Section 3.5.

The selected parameter settings for the first iteration included a polynomial kernel and cost parameter of 1. Those selected for the second included a linear kernel with a cost parameter of 1. These settings were found to produce the most reasonable results in terms of the number of matches. Other parameter setting performed poorly, sometimes resulting in more matches than the minimum number possible.

Manual review of the match sets produced by each iteration revealed several interesting things. In iteration 1, the classifier only matched incidents that were no more than 5 hours apart. Although comparisons were allowed over a 24 hour period, incidents were classified as matches within a similar time window to that identified by our previous approach. One possible explanation for this is the choice of timestamp comparison. Because the scores assigned when comparing two timestamps decays linearly over 24 hours, it is likely that this influenced the classifiers behaviour in preference of a smaller time window. Indeed, the features used in the comparison step are known to play a significant role in shaping the matching outcome, regardless of the classification

method used.

In addition, the choice of SVM parameters greatly impacted the classifier's performance, with some settings emphasising postcode comparisons over temporal ones. For example, one parameter variation returned matches solely based on an exact match on postcode sector, even in cases where incidents had no temporal relation.

In iteration 2, we changed the way we compared postcode information. Instead of exact comparisons between postcode sectors, the percentage overlap between postcode units and sectors was used to quantify agreement. As before, we searched for the inclusion of additional matches that occurred on different sides of a postcode sector boundary. Surprisingly, this change had little impact on the resulting match sets. In fact, no incidents involving different postcode sectors were included in the match sets across all reasonable parameter variations.

One reason for this is that the number of incidents that occur close to a postcode boundary were found to be very few relative to those that did not. Specifically, the change in the postcode comparison introduced only 59 additional potential matches on top of the 733 that already matched exactly on postcode sector. Further scrutiny of the selected training examples also revealed that no cross-boundary examples were included in the seed training set. These factors likely hindered the classifier's ability to learn distinctions between matches and non-matches involving different postcodes. The final iteration of this approach therefore used an exact match on postcode sector and a linear extrapolation between exact similarity and exact dissimilarity over a 24 hour period.

Overall, the iterative SVM was able to identify decision criteria for classifying record pairs that compares to the criteria identified using a deterministic approach. However, due to the way training data is generated, the SVM struggled to learn more nuanced agreement patterns that were otherwise captured using the deterministic approach.

In the context of streamlining the linkage process, the SVM did not perform better than the deterministic approach. This is because it required fine-tuning of a variety of parameters and careful design of the comparison vectors based on detailed domain knowledge to achieve optimal performance. The large number of possible comparison and parameter combinations makes evaluation a difficult and time consuming process, especially if manual review is necessary. On the other hand, evaluation of the deterministic approach proved far more straightforward. It is for this reason that we use the results from the deterministic approach in further analyses.

Note, however, that our evaluation of both approaches is limited due a lack of detailed incident information and ground truth data. Fully assessing the performance of each approach would require extensive review of incident information provided by both police and ambulance services. While this was not possible here, we hope that our work can serve as a guide for such future endeavours.

4.3 Extent of Overlap

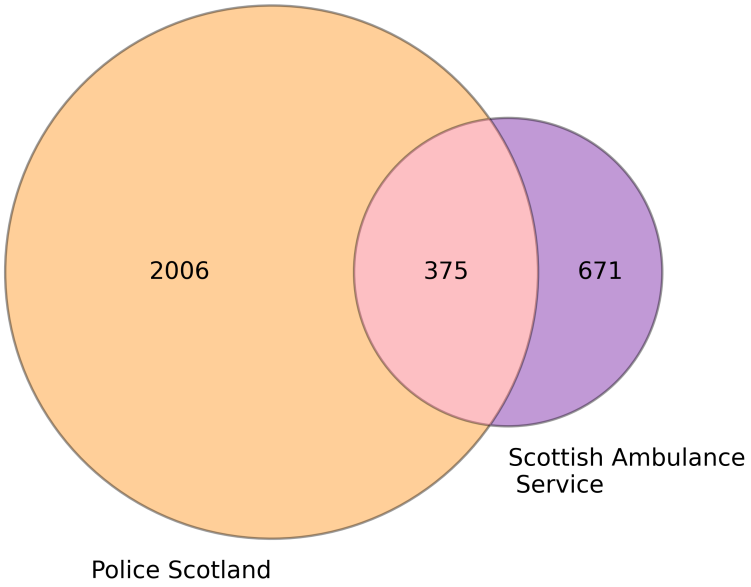

Figure 4.2: Venn diagram depicting the overlap between police and ambulance datasets.

Previous research into police and ambulance data matching has yielded a small number of findings as to the extent of data overlap. One study examined hotspots of community violence-related calls to police and ambulance services, finding that the respective hotspots of violence overlapped by 50% on average. It was also reported that only 62% of incidents in the ambulance data were present in the police data [1]. Another study analysed the potential value of ambulance data for violence prevention using datasets provided by West Midlands Police and the West Midlands Ambulance Service. It was

found that between 66 and 90% of incidents recorded by ambulance services were not found in police data [35]. Similarly to ours, their approach looked at spatial and temporal overlap between the two datasets.

Using our deterministic approach to linkage, we found that approximately 56% of knife-related injuries recorded by the Scottish Ambulance Service were also recorded by Police Scotland, as depicted in Figure 4.2. In other words, 44% of knife-related injuries were not reported to the police. In line with previous work, this suggests that ambulance records may contain substantial new information relating to knife crime and highlights the importance of data sharing across police and public health services.

4.4 Temporal Trends

The number of incidents in each dataset were compared by time of day as depicted in Figure 4.3. Notably, the overall distribution of incidents exhibits remarkable similarity between the police and ambulance datasets, with minimal reports made in the morning and pronounced peaks during night time hours. This pattern suggests that individuals are just as likely to contact police and ambulance services, regardless of the hour. Therefore, underreporting of incidents is unlikely to be influenced by time of day.

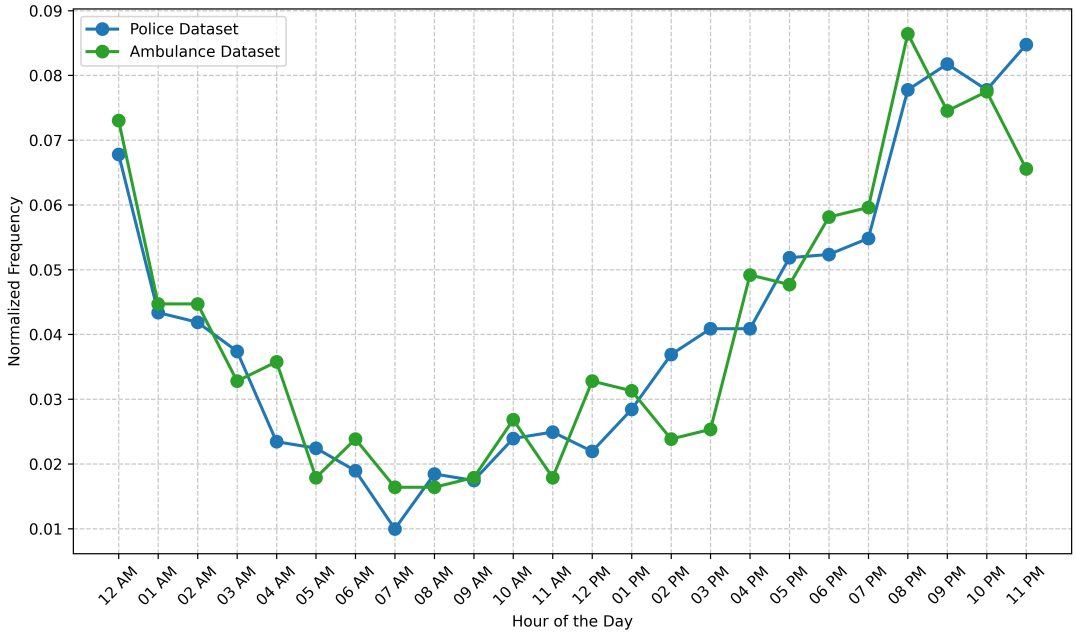

Figure 4.3: The normalised incident frequencies in police and ambulance datasets by time of day.

In addition, the frequency of incidents occurring by day of the week is shown in Figure 4.4. Once again, both datasets exhibit similar trends, with fewer incidents occurring during the weekdays as compared to weekends. We note that the ambulance dataset has a slightly higher incident frequencies on Wednesday and Saturday, potentially suggesting underreporting on those days of the week. However, given the limited number of incidents in the ambulance dataset, it remains uncertain whether this observation stems from chance fluctuations. Conducting an analysis over an extended timeframe would offer more reliable insights. This was similarly found to be the case for the monthly incident distributions, which is why we chose not to include the results here (see Appendix A.1).

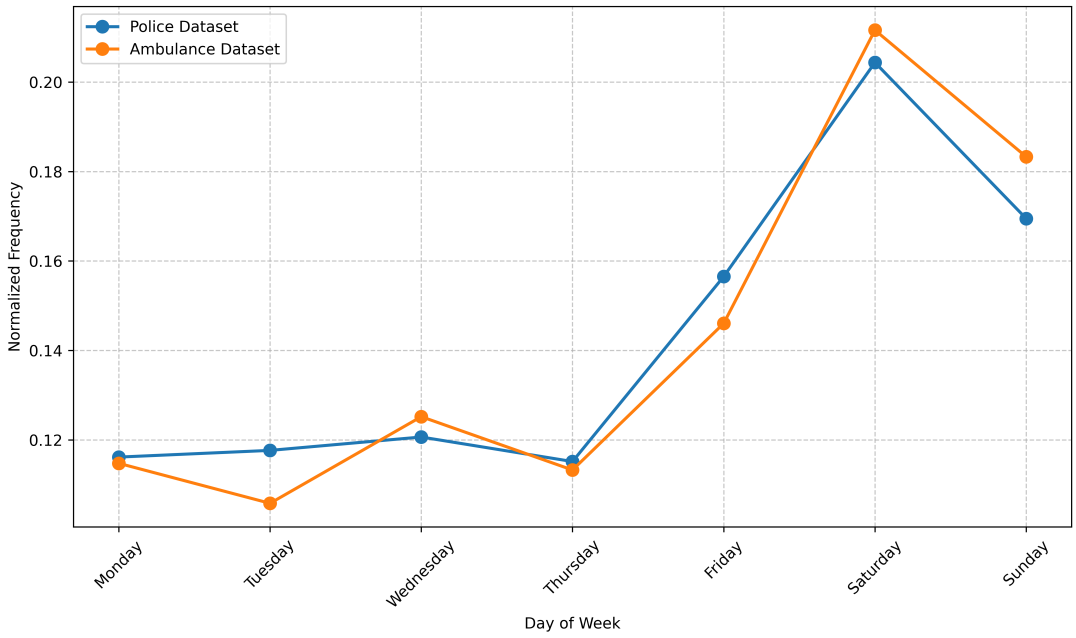

Figure 4.4: The normalised incident frequencies in police and ambulance datasets by day of the week.

4.5 Geographical Coverage

Examining the geographical allocation of reported incidents revealed several interesting things. First, the police and ambulance datasets covered 567 and 349 unique postcode sectors, respectively. Among these sectors, 327 were common to both datasets, leaving 22 unique to the ambulance dataset and 240 unique to the police dataset. One reason for this discrepancy could be variations in data collection and reporting. Another may be variations in the perception and severity of injuries, which differ from

incident to incident. For instance, some incidents may not require an ambulance following a call to the police.

Postcode Area	Proportion of Unreported (%)
Edinburgh	14.9
Glasgow	14.0
Motherwell	12.3
Kirkcaldy	9.5
Dundee	9.4
Aberdeen	7.4
Kilmarnock	7.3

Table 4.1: Proportions of unreported incidents by postcode area.

Furthermore, densely populated areas have much higher incident frequencies. This aligns with our expectations as a large number of people in close proximity often correlates with a greater number of incidents. In general, higher density areas were found to have slightly higher rates of underreporting, as shown in Table 4.1. To account for the small number of incidents in the datasets, the proportion of incidents that were not reported to the police are calculated for each postcode area. Postcode areas with very few reported incidents were also omitted from the analysis to ensure statistical validity.

Figure 4.5 illustrates the frequencies of incidents per postcode sector that were not captured by police records in the Glasgow area. Lighter regions denote higher incident numbers, providing an overview of potential areas of underreporting. We note, however, that analysis at the postcode sector level may be less informative due to large geographical coverage, especially in rural areas. Future efforts could benefit from more detailed geographical analysis, enabling closer examination of underreporting patterns.

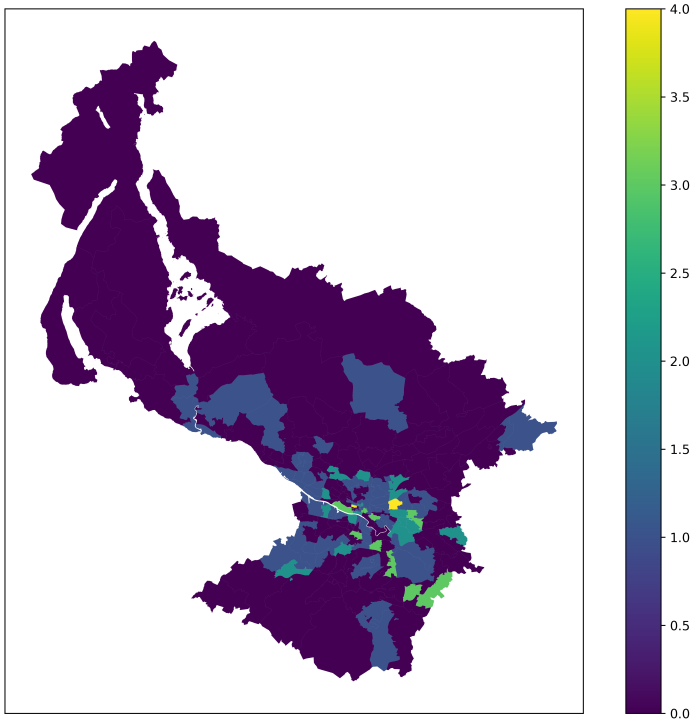

Figure 4.5: Heatmap depicting the frequency of incidents that were not found in police records in each postcode sector in the Glasgow postcode area.

Chapter 5

Conclusion

The purpose of this project was to determine suitable matching criteria for linking police and ambulance datasets and, through this, understand the extent of overlap in incidents recorded in the two datasets. Another aim that emerged was the automation of matching criteria. The two primary reasons for doing this were first, to streamline the record linkage process, and second, to identify potential patterns of underreporting.

Two classification approaches were used: deterministic and iterative SVM. The first approach allowed the design of a simple set of matching criteria to binary classify records into matches and non-matches. It was identified that a 4 to 5 hour window was the most suitable time frame for striking a balance between the number of true matches and the number of false matches. This aligns with findings in previous work [35]. An approach for matching postcode information using the percentage overlap between postcode units and postcode sectors was used in combination with the selected time window. This approach required that incidents occurring more than one hour apart, but in different postcode sectors, were excluded from the match set. Interestingly, seven of the twelve matches that were included as a result were true positives, linking related incidents that occurred on opposite sides of a postcode sector boundary.

The second approach used a nearest based method to generate seed training examples, which in turn were used to iteratively train an SVM classifier. This method yielded similar matching criteria to those identified using the deterministic approach, linking incidents that occurred no more than 5 hours apart. However, the SVM classifier was unable to capture additional agreement patterns relating to incidents at postcode sector boundaries. This was likely due to the small number of boundary cases included in the data and, by extension, the training seeds.

For the purpose of streamlining, deterministic classification outperformed iterative

SVM classification in the presence of two linkage fields and no ground truth data. The primary limitation of the SVM was that it required careful fine tuning of a number of various parameters. This made evaluation a difficult task when having to manually review results. Evaluation of the deterministic approach, on the other hand, proved far simpler. This method enabled a pragmatic approach to balancing precision and recall, underlining its suitability for use in resource-limited environments. For this reason, the deterministic approach was carried forward for further analyses. We note, however, that SVM classification has greater potential for scalability in projects with a larger number of linkage fields.

The overlap between police and ambulance datasets revealed that ambulance records contain substantial new information related to incidents of knife crime, with approximately 44% of knife-related injuries not reported to the police. This suggests that police are not aware of a large number of stabbings that are otherwise recorded by ambulances services, leaving a gap in the understanding of knife crime in Scotland. These results, consistent with prior work that cuts across police and public health care, highlight the benefits of continued interagency data sharing and integration.

While we recognise that ambulance data has value for violence prevention, our analysis was limited due to a lack of detailed incident information. A full understanding of the extent and reasons behind gaps in reporting will require further collaborative investigation between police and public health services. By harnessing the strengths of different agencies, a better understanding of violent crime is possible. The need for more research and development of such methods therefore seems clearly justified.

Bibliography

[1] Barak Ariel, Cristobal Weinborn, and Adrian Boyle. Can routinely collected ambulance data about assaults contribute to reduction in community violence? *Emergency medicine journal*, 32(4):308–313, 2015.

[2] Rohan A. Baxter, Peter Christen, and Tim Churches. A comparison of fast blocking methods for record linkage. In *Knowledge Discovery and Data Mining*, 2003.

[3] Mikhail Bilenko, Beena Kamath, and Raymond J Mooney. Adaptive blocking: Learning to scale up record linkage. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 87–96. IEEE, 2006.

[4] Adrian A Boyle, Katrina Snelling, Laura White, Barak Ariel, and Lawrence Ashelford. External validation of the cardiff model of information sharing to reduce community violence: natural experiment. *Emergency medicine journal*, 30(12):1020–1023, 2013.

[5] Li Cai and Yangyong Zhu. The challenges of data quality and data quality assessment in the big data era. *Data science journal*, 14:2–2, 2015.

[6] Peter Christen. A two-step classification approach to unsupervised record linkage. In *Proceedings of the sixth Australasian conference on Data mining and analytics-Volume 70*, pages 111–119, 2007.

[7] Peter Christen. Automatic training example selection for scalable unsupervised record linkage. In *Advances in Knowledge Discovery and Data Mining: 12th Pacific-Asia Conference, PAKDD 2008 Osaka, Japan, May 20-23, 2008 Proceedings 12*, pages 511–518. Springer, 2008.

[8] Peter Christen. A survey of indexing techniques for scalable record linkage and deduplication. *IEEE transactions on knowledge and data engineering*, 24(9):1537–1555, 2011.

[9] Peter Christen. *Data Matching*. Springer Berlin, Heidelberg, 2012.

[10] David E Clark. Practical introduction to record linkage for injury research. *Injury Prevention*, 10(3):186–191, 2004.

[11] Karen Critchley and Zara Quigg. A cross-sectional study of child injury ambulance call-out characteristics and their utility in surveillance. *Journal of Paramedic Practice*, 11(7):282–292, 2019.

[12] James C Doidge and Katie Harron. Demystifying probabilistic linkage: Common myths and misconceptions. *International journal of population data science*, 3(1), 2018.

[13] M.G. Elfeky, V.S. Verykios, and A.K. Elmagarmid. Tailor: a record linkage toolbox. In *Proceedings 18th International Conference on Data Engineering*, pages 17–28, 2002.

[14] Ivan P Fellegi and Alan B Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

[15] Curtis Florence, Jonathan Shepherd, Iain Brennan, and Thomas Simon. Effectiveness of anonymised information sharing and use in health service, police, and local government partnership for preventing violence related injury: experimental study and time series analysis. *BMJ*, 342, 2011.

[16] Carol Friedman and Robert Sideli. Tolerating spelling errors during patient validation. *Computers and Biomedical Research*, 25(5):486–509, 1992.

[17] Benjamin J Gray, Emma R Barton, Alisha R Davies, Sara J Long, Janine Roderick, and Mark A Bellis. A shared data approach more accurately represents the rates and patterns of violence with injury assaults. *J Epidemiol Community Health*, 71(12):1218–1224, 2017.

[18] Lifang Gu and Rohan Baxter. Decision models for record linkage. *Data Mining: Theory, Methodology, Techniques, and Applications*, pages 146–160, 2006.

[19] Mauricio A Hernández and Salvatore J Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data mining and knowledge discovery*, 2:9–37, 1998.

[20] A Howe and M Crilly. Identification and characteristics of victims of violence identified by emergency physicians, triage nurses, and the police. *Injury Prevention*, 8(4):321–323, 2002.

[21] Anna Jurek, Jun Hong, Yuan Chi, and Weiru Liu. A novel ensemble learning approach to unsupervised record linkage. *Information Systems*, 71:40–54, 2017.

[22] Mayank Kejriwal and Daniel P Miranker. Semi-supervised instance matching using boosted classifiers. In *The Semantic Web. Latest Advances and New Domains: 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31–June 4, 2015. Proceedings 12*, pages 388–402. Springer, 2015.

[23] Heikki Keskustalo, Ari Pirkola, Kari Visala, Erkka Leppänen, and Kalervo Järvelin. Non-adjacent digrams improve matching of cross-lingual spelling variants. In *String Processing and Information Retrieval: 10th International Symposium, SPIRE 2003, Manaus, Brazil, October 8-10, 2003. Proceedings 10*, pages 252–265. Springer, 2003.

[24] Michael D Larsen and Donald B Rubin. Iterative automated record linkage using mixture models. *Journal of the American Statistical Association*, 96(453):32–41, 2001.

[25] Marie-Jeanne Lesot, Maria Rifqi, and Hamid Benhadda. Similarity measures for binary and numerical data: a survey. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 1(1):63–84, 2009.

[26] Un Yong Nahm, Mikhail Bilenko, and Raymond J Mooney. Two approaches to handling noisy variation in text mining. In *Proceedings of the ICML-2002 workshop on text learning (TextML'2002)*, pages 18–27. Citeseer, 2002.

[27] Gonzalo Navarro. A guided tour to approximate string matching. *ACM computing surveys (CSUR)*, 33(1):31–88, 2001.

[28] Howard B Newcombe, James M Kennedy, SJ Axford, and Allison P James. Automatic linkage of vital records: Computers can be used to extract" follow-up" statistics of families from files of routine records. *Science*, 130(3381):954–959, 1959.

[29] Toan C Ong, Michael V Mannino, Lisa M Schilling, and Michael G Kahn. Improving record linkage performance in the presence of missing linkage data. *Journal of biomedical informatics*, 52:43–54, 2014.

[30] George Papadakis, Dimitrios Skoutas, Emmanouil Thanos, and Themis Palpanas. Blocking and filtering techniques for entity resolution: A survey. *ACM Computing Surveys (CSUR)*, 53(2):1–42, 2020.

[31] Leo L Pipino, Yang W Lee, and Richard Y Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.

[32] Banda Ramadan and Peter Christen. Unsupervised blocking key selection for real-time entity resolution. In *Advances in Knowledge Discovery and Data Mining: 19th Pacific-Asia Conference, PAKDD 2015, Ho Chi Minh City, Vietnam, May 19-22, 2015, Proceedings, Part II 19*, pages 574–585. Springer, 2015.

[33] Sunita Sarawagi and Anuradha Bhamidipaty. Interactive deduplication using active learning. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–278, 2002.

[34] Adrian Sayers, Yoav Ben-Shlomo, Ashley W Blom, and Fiona Steele. Probabilistic record linkage. *International journal of epidemiology*, 45(3):954–964, 2016.

[35] Alex Sutherland, Lucy Strang, Martin Stepanek, Chris Giacomantonio, Adrian Boyle, and Heather Strang. Tracking violent crime with ambulance data: how much crime goes uncounted? *Cambridge Journal of Evidence-Based Policing*, 5(1-2):20–39, 2021.

[36] M Teff. Information sharing to tackle violence. guidance for community safety partnerships on engaging with the nhs, 2012.

[37] Vassilios S Verykios, Ahmed K Elmagarmid, and Elias N Houstis. Automating the approximate record-matching process. *Information sciences*, 126(1-4):83–98, 2000.

[38] Alison Warburton and Jonathan Shepherd. Development, utilisation, and importance of accident and emergency department derived assault data in violence management. *Emergency medicine journal : EMJ*, 21:473–7, 08 2004.

[39] William E Winkler. *Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage*. US Bureau of the Census Washington, DC, 2000.

[40] Huiping Xu, Xiaochun Li, Changyu Shen, Siu L Hui, and Shaun Grannis. Incorporating conditional dependence in latent class models for probabilistic record linkage. *The Annals of Applied Statistics*, 13(3):1753–1790, 2019.

Appendix A

A.1 Incident Frequencies by Month

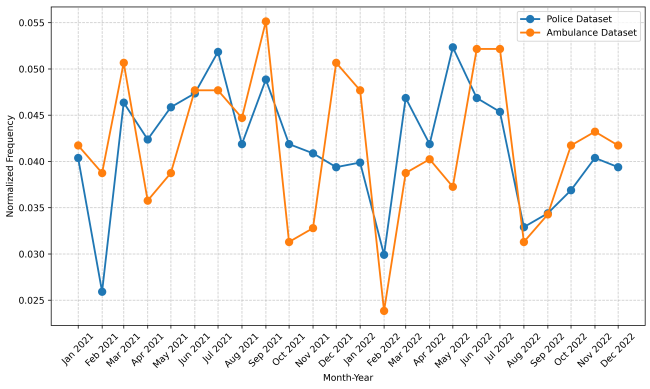

Figure A.1: The normalised incident frequencies in police and ambulance datasets by time of day.

Appendix B

B.1 Geographical Coverage of Police Dataset

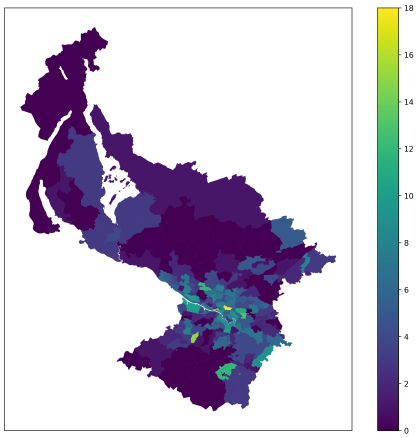

Figure B.1: Heatmap depicting the frequency of incidents in the police dataset for each postcode sector in the Glasgow area.

B.2 Geographical Coverage of Ambulance Dataset

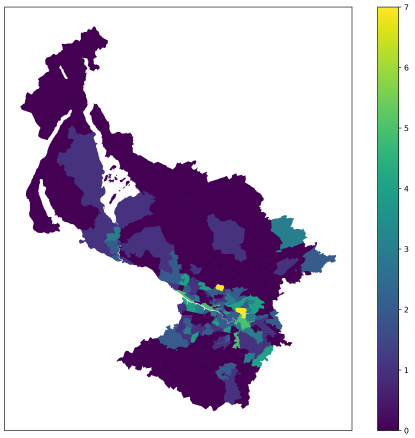

Figure B.2: Heatmap depicting the frequency of incidents in the ambulance dataset for each postcode sector in the Glasgow area.