Improving the Accuracy of a Sequential Mining Model for Hurricane Trajectory Prediction

Jasmin Bissonnette, Caden Marofke, Taylor Cox

Department of Computer Science
University of Manitoba
Winnipeg, Manitoba
Email: {bissonnj, marofkec, coxt3}@myumanitoba.ca

Abstract—In the 21st century alone, hurricanes have caused billions of dollars worth of damage due to property destruction, personal danger, and civil unrest. The catastrophic impact of hurricanes and other tropical storms has motivated extensive research into hurricane forecasting. Hurricane trajectory prediction is a subfield of hurricane forecasting that traditionally uses meteorological analysis to determine the future path of a given hurricane. While meteorological methods of hurricane trajectory prediction have been effective in the past, little attention has been paid to hurricane trajectory prediction with respect to historical frequent pattern mining. In this work, we propose multiple dimensions for improving the accuracy of data mining-based hurricane trajectory prediction. We propose the use of Haversine distance, a real-world distance formula for reliable hurricane pattern matching accuracy. We also propose the use of weighted data to favor recent hurricane trends when training the trajectory prediction model. Finally, we use training data exclusively from the years 1950 to 2000 to reduce the impact of historical inaccuracies in hurricane recording methodologies. After implementing all three improvement dimensions on a contemporary data mining model for hurricane trajectory prediction based on AprioriAll, results show a prediction-correctness ratio up to 10.0% higher than the current state-of-the-art, at 75.0%.

Index Terms—Hurricane Trajectory, AprioriAll, Prediction, Pattern Matching

I. INTRODUCTION

Hurricanes are a specific type of tropical cyclone exclusive to the Atlantic basin. Hurricanes are divided into five categories according the Saffir-Simpson scale, based on wind-speed and expected damage. In the United States alone, hurricanes incur damages which can reach $157BN [1]. The damages incurred by hurricanes do not only include property damage, but also personal danger and in some cases widespread civil unrest [2]. These damages motivate extensive research in the field of hurricane prediction analysis. Hurricane prediction analysis is traditionally divided into the subfields of hurricane intensity prediction and hurricane trajectory prediction. This paper focuses exclusively on the subfield of hurricane trajectory prediction. In particular, the goal of this paper is to investigate and refine the existing-state-of-the art in hurricane trajectory prediction systems based on sequential data mining.

Data mining is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [3]. Data mining applies to an extremely wide range of applications, including clickstream mining, social network community detection, and artificial intelligence. The data mining subdiscipline of particular interest in this work is *sequential* data mining. Sequential data mining uses a database of chronologically ordered inputs to mine frequent patterns and interesting association rules. These association rules are used to make predictions on future data entries such as future customer transactions. More specifically, association rules generated by sequential data mining algorithms including AprioriAll, SPIRIT and PrefixSpan can be used in the field of hurricane trajectory prediction.

To the knowledge of the authors, the existing standard in hurricane trajectory prediction via data mining is developed by Dong et al in [4]. The work cited proposes a hurricane trajectory prediction model based on data mining (HTPDM) that uses a modified variant of AprioriAll, a common sequential mining algorithm [6]. HTPDM executes AprioriAll against historical hurricane trajectories from 1900 to 2000 to generate interesting trajec-

tory association rules. These association rules are compared against test data composed of hurricane trajectories from 2001 to 2008. HTPDM divides a given testing trajectory into two parts: the initial trajectory T_i and the terminal trajectory T_t. T_i is compared against all association rule antecedents until a sufficiently matching rule R is found based on a predetermined fitness function. The consequent of R is then compared against T_t using pattern-matching techniques to determine if the prediction was correct. The results of HTPM showed that the model was able to achieve a best-case correctness ratio of 65.0%. In this work, HTPDM will be developed and various strategies will be employed to increase the correctness ratio. Reducing the impact of historical factors, favoring the impact of recent trends, and improving the realistic nature of the hurricane trajectory prediction model result in a new model dubbed WARD-HTP (Weighted-Asset, Realistic-Distance Hurricane Trajectory Predictor). The strategies employed in WARD-HTP ultimately result in a correctness ratio of 75.0%, a 10% improvement over HTPDM.

The remainder of this paper is structured as follows: First, Section II gives a detailed survey of related work in the hurricane prediction and sequential mining disciplines. Next, section III will cover the general motivations of this work with respect to hurricane trajectory prediction, and the specific motivations of this work with respect to improving the current state-of-the-art. Section IV will provide a description of the problem space, which includes specific implementation considerations in hurricane trajectory prediction. Section V will give an exposition of the solutions proposed in WARD-HTP for increasing the correctness ratio in hurricane trajectory prediction. Following that, Section VI will detail the experimental setup, including the training data and other experimental variables. Section VII will evaluate the results of the experiments, which show that the ideal configurations of WARD-HTP result in a maximal correctness ratio of 75.0%. This work will then conclude with section VIII, capturing the conclusions and directions of future work.

II. RELATED WORK

The development of an accurate hurricane trajectory prediction model is a problem which has received little attention in the data mining literature. Works including [7] and [8] have been completed to develop data-mining approaches to hurricane intensity prediction. These works however do not describe approaches to hurricane trajectory prediction. Hurricane trajectory prediction has been covered in previous meteorological and geographic information system (GIS) works. These types of hurricane trajectory forecasting models include those seen in [9] or [10]. The intersection between data mining and hurricane trajectory prediction has been sparsely investigated in the literature. Publications such as [11] proposed a clustering-based model for typhoon track prediction based on specific characteristics of historical data. Typhoons are then clustered into groups depending on the general trajectory patterns they exhibit. While the described work provided novel results in typhoon track clustering, it is not clear if the proposed solution is able to predict the trajectories of new storms. Additionally, the clusters developed apply specific to typhoons, which are waterborne storms native to the east-pacific seaboard. Such storm patterns may not be applicable to patterns found in hurricanes, which are waterborne storms of the west-atlantic basin. For this reason, the proposed clustering patterns were not used in this work.

The primary baseline of this work is [4], which uses a specific implementation of AprioriAll (first proposed in [6]) to determine frequent hurricane trajectory patterns. These trajectory patterns correspond to interesting trajectory association rules, which are used to train a prediction model. The implementation of AprioriAll used in [4] is also used as the foundational model for this work. AprioriAll transforms a transaction database into a *customer* database, where each entry contains a customer and an ordered sequence of purchased items. The ordered sequence database is then mined for frequent ordered patterns, or frequent sequences. The modification of Apriori proposed in [4] mines frequent ordered patterns and thus skips the transformation step of AprioriAll.

The model proposed in this work uses historical hurricane trajectory readings from HURDAT, the canonical hurricane database maintained by the National Hurricane Center [5]. The data in HURDAT2 is ordered into $\langle Hurricane, Trajectory \rangle$

pairs, omitting the need for the sequence mapping step required at initialization of AprioriAll. The Atlantic hurricane trajectories recorded in HURDAT2 span a very large distance, some of which reaching over 10,000 kilometers in total distance travelled. As a result, the coordinates recorded are limited to those within a range of 0 to 50 degrees north, and 20 to 100 degrees west. This degree range is approximately the size of the south Atlantic basin.

III. MOTIVATION

There are two categories of motivation behind this work. The first motivation of this work is to increase the total investigation completed in the hurricane trajectory prediction subdiscipline. Little research effort has been allocated to this particular problem space of high economic, and social impact. The authors expect this work to contribute to the growing effort by academic and government researchers to develop reliable methods of hurricane trajectory prediction in the interest of public safety and economic stability.

The second motivation behind this work is the specific goal of improving the correctness ratio in the current standard data mining model for hurricane trajectory prediction. As previously described, HTPDM had a best case trajectory prediction correctness ratio of 65.0%. This means that of all testing trajectory for a which a matching rule was successfully produced, 65.0% of such matches were correct predictions. When HTPDM accounts for testing trajectories for which no rule could be matched, the correctness ratio decreases by 7.5% to a final value of 57.5% correctness. This decrease in accuracy is expected as some hurricane trajectories are outliers which cannot be predicted by frequent pattern mining, as will be discussed further in this work.

The motivation of improving the correctness ratio in the model encompasses the challenges of improving the best-case correctness ratio (the correctness ratio when ignoring rule-miss incidents) and the worst-case correctness ratio (the correctness ratio when accounting for rule-miss incidents). To resolve the motivations of this work, a series of improvement strategies are proposed as applications to the current hurricane trajectory prediction model based on AprioriAll. In general, these strategies seek to improve the realism of the prediction model, reduce historical factors which may negatively affect the model's correctness reatio, and favor recent hurricane trajectory trends which are likely to be reflected in current and future hurricane patterns.

IV. PROBLEM DESCRIPTION

In this section, the problem of achieving accurate hurricane trajectory prediction is explained in detail. This problem consists of multiple challenges which apply generally to sequential data mining, and specifically to the problem domain of hurricane location analysis. Many of the domain-specific challenges of hurricane trajectory prediction arise from the fact that hurricane points are tracked as a series of $\langle Latitude, Longtitude \rangle$ points, which do *not* directly correspond to traditional (x, y) coordinates, as latitude and longtitude are points imposed upon a sphere. Challenges including accurate distance evaluation are results of the transition from a Euclidean to a Polar coordinate system.

The exposition of the hurricane prediction problem domain is divided into three parts. The first aspect of the problem domain concerns the underlying sequential mining approach used. This includes a thorough description of AprioriAll, its intended usage, and the reaction of AprioriAll to its input parameters. Next, the subproblem of region discretization will be discussed. Since $\langle Latitude, Longtitude \rangle$ points constitute continuous space, hurricane trajectories are converted to sequences of discrete regions to enable the discovery of frequent patterns. Finally, problems involving hurricane trajectory matching will be discussed. These problems include pattern matching for determining best fit rules and pattern matching for determining trajectory prediction correctness.

A. AprioriAll

In HTPDM and WARD-HTP, the algorithm that generates trajectory prediction possibilities is based on AprioriAll. AprioriAll was first proposed at IBM research as a method for discovering frequent sequential patterns in transactional databases [6]. In its original implementation, AprioriAll is able to generate assocation rules describing sentences of the form "Customers who buy P may also buy Q". In the hurricane trajectory prediction problem,

AprioriAll is used to generate association ruls describing sentences of the form "Hurricanes located at P may later be located at Q". In the context of AprioriAll, hurricanes are considered $customers$ and their sequences of recorded points are considered $transactions$. Consider the high-level implementation steps of AprioriAll as follows:

1) Map the transaction database into a customer database.
2) Generate frequent sequences based on minimum support.
3) Generate association rules based on minimum confidence.

1) Mapping: In step (1) of AprioriAll, a given transaction database D is grouped by customers, and the frequent items purchased by customers in individual transcations are recorded in a temporary table. The contents of this temporary table are mapped to corresponding sequences to enable the mining of frequent transaction sequences from the original database. The intention of step (1) is to map D in its original $\langle TransactionID, CustomerID, Items \rangle$ database form into a database D' of the form $\langle CustomerID, ItemSequence \rangle$.

The hurricane trajectory model in this work does not implement step (1), as the training data from HURDAT2 was instead translated into a $\langle HurricaneID, CoordinateSequence \rangle$ database as a preprocessing step before any aspects of the model are executed. By directly translating the data instead of processing it with AprioriAll, all trajectories in the original database are preserved. This includes coordinates which may ultimately be infrequent. The sequential mapping of AprioriAll is performed as a preprocessing step so that the database used in the prediction model will always consist of hurricanes and their coordinate sequences. This allows the prepared (mapped) database to be mined directly without the need to complete the mapping step every time AprioriAll is executed.

Figure (1) shows an example of how the trajectory database is structured after the mapping preprocessing step. Only two dimensions are needed in the database, being the hurricane ID and the sequence of latitude-longtitude coordinates. All hurricanes in HURDAT2 have an ID corresponding to the location (i.e. the Atlantic), the index ordered

Fig. 1. Sample of mapped Hurricane-Coordinates database

HurricaneID	LatLonSequence
AL081976	⟨ (26.0:-84.0),(25.3:-83.3),(24.7:-82.7) ⟩
AL091976	⟨ (31.7:-68.2),(33.4:-67.5),(35.2:-66.4) ⟩
AL101976	⟨ (14.0:-48.0),(14.0:-49.3),(14.0:-50.6) ⟩
AL111976	⟨ (12.5:-37.5),(13.0:-39.0),(13.5:-40.5) ⟩

by appearance in the hurricane year, and the year itself. For example, hurricane Katrina of year 2005 corresponds to AL122005. The preprocessing step proposed is executed exactly once, allowing the hurricane database to be mined without the initial grouping and mapping stage.

2) Sequence Generation: Section (2) of AprioriAll initiates the generation of frequent sequences. In the case of hurricane trajectory prediction, these sequences correspond to frequent recorded hurricane coordinate sequences. Frequent hurricane coordinate sequences are those which occur at least a minimum number of times, as defined by the *minimum support*, or minsup. In general terms, the sequence generation step of AprioriAll develops frequent ordered patterns of increasing length, beginning with frequent single-item sequences (corresponding to frequent single-point sequences). This step begins by scanning the $\langle HurricaneID, CoordinateSequence \rangle$ database for frequent single-point sequences. The frequent single-point trajectories are then converted to candidate two-point trajectories as the result of an inner join. The candidate two-point trajectories are scanned against the database to determine which candidates satisfy the minimum support. Another inner join generates the candidate 3-length trajectories. Once the sequence generation level reaches three, each candidate sequence must be pruned if any of their candidate subsequences are infrequent. Continuous execution of AprioriAll until completion results in a collection of frequent hurricane trajectory patterns, as defined by the minumum support.

Algorithm 1 Modified AprioriAll: No Litemset phase

1: $L_1 \leftarrow \{Frequent\ Single\ Points\}$
2: **for** $(k = 2; L_{k-1} \neq \emptyset; k++)$ **do**
3: $\quad C_k \leftarrow Candidates\ from\ Inner\ Join$
4: $\quad Prune\ C_k\ of\ infrequent\ subsequences$
5: $\quad$ **for** $sequence\ S \in HurricaneDB$ **do**
6: $\quad\quad Increment\ c \in C_k\ support\ if\ c \in S$
7: $\quad$ **end for**
8: $\quad L_k \leftarrow \{c \in C_k|\ sup(c) \geq minsup\}$
9: **end for**
10: $FrequentTrajectories \leftarrow All(L_k)$

Along with HTPDM, WARD-HTP executes AprioriAll with a slight modification to the existing algorithm proposed in [6]. Whereas L1, the frequent single-item sequences is generated by the L-Itemset (mapping) phase of AprioriAll, the variant of AprioriAll used in hurricane trajectory prediction skips the L-Itemset phase, and thus scans the $\langle HurricaneID, CoordinateSequence\rangle$ directly for frequent single-point trajectories. Otherwise, the remainder of AprioriAll is used in this work as originally designed. AprioriAll is used to ensure that the order of recorded hurricane coordinate sequences is preserved. This is a product of the inner-join step. Consider two hurricane trajectories T_1 and T_2. T_1 and T_2 form a new trajectory T_3 if and only if the L-1 terminal coordinates of T_1 correspond to the L-1 initial coordinates of T_2, where L is the current level of AprioriAll being executed. For example, if L=2, then AprioriAll scans for frequent 2-coordinate sequences, and uses those frequent 2-coordinate sequences to generate 3-coordinate sequences based on trajectory matching. If the matching condition is satisfied, a new trajectory is generated. The inner-join step of AprioriAll emulates the required hurricane trajectory-matching to develop frequent hurricane trajectory patterns. Once all possible frequent hurricane trajectory patterns are discovered, the second step of AprioriAll completes.

3) Rule Generation: The final step of AprioriAll is the rule-generation step. Though association rule generation is not a specific stage of the AprioriAll algorithm, it is traditionally the final stage of any association rule mining approach. The rule generation step is responsible for developing interesting association rules. As mentioned, rules generated by the hurricane trajectory prediction model proposed will satisfy sentences of the form "Hurricanes located at P may later be located at Q".

Association rules are derived directly from frequent patterns generated by AprioriAll or any modified versions. These rules consist of four dimensions: the antecedent, consequent, support and confidence. An association rule R is generated by partitioning a frequent sequence S into two congruent parts. The left partition of S constitutes the antecedent of R, and the latter partition is the consequent of R. The support of R is the original support of S. The confidence of R is the support of S divided by the support of the subsequence given by the former partition.

Rules with confidence of at least the *minimum confidence* (minconf) are stored in a table of interesting association rules. The rules generated by AprioriAll are the candidates for hurricane trajectory pattern matching. Hurricane trajectories are matched against all association rule antecedents. The consequent of the closest match rule is then used to predict the hurricane terminal trajectory. If the association rule consequent matches the actual hurricane terminal trajectory, the prediction is considered correct. Without the association rules formed by AprioriAll and the rule generation step, data mining-based hurricane trajectory prediction cannot be executed. HTPDM and WARD-HTP both use historically frequent hurricane trajectory to predict the result of future hurricane paths.
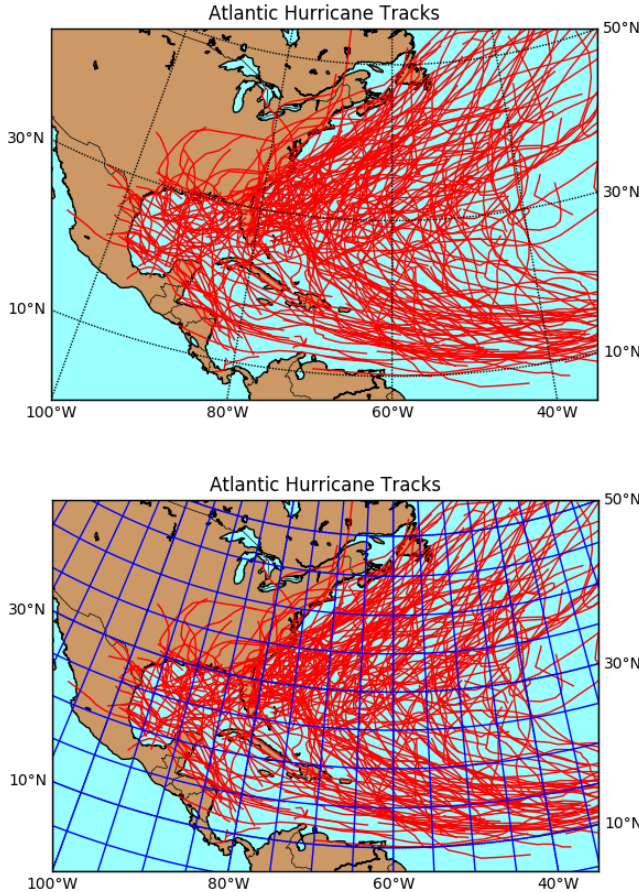
B. Region Discretization

This subsection covers problems related to the discretization of $\langle Latitude, Longtitude\rangle$ points. As shown in figure (1), the raw recorded hurricane coordinates are expressed in degrees and minutes. The excessively fine-grained nature of the raw data prevents the data from being able to produce meaningful frequent coordinate sequences. Therefore, effort must be taken to discretize the coordinate system, allowing more coarse-grained frequent patterns to be produced.

In HTPDM and WARD-HTP, recorded hurricane coordinates are discretized into approximately

square blocks of a specific size, expressed in degrees. Consider the first row of coordinates shown in figure (1): (12.5:-37.5), (13.0:-39.0), (13.5:-40.5). Mapping this coordinate sequence to a discritization block of size 5, the new trajectory becomes (2:-7), (2:-7), (2:-8). Duplicate coordinates are not counted, so the trajectory ultimately reduces to (2:-7), (2:-8). Note that the degree-size of the discretization block is expressed as an integer, which means that all coordinates are mapped to their lowest divisor relative to the discretization size. This enables the coarse-grained mapping of coordinates to their corresponding region block.

Fig. 2. Sample Hurricane Trajectory Paths

Consider figure (2), generated using a sample of the recrded hurricane trajectories in HURDAT2 with the visualization tool Matplotlib [12]. As shown in the pair of images, region discretization organizes continuous hurricane coordinate sequences into discrete region sequences of approximately uniform size and square shape. A discretization block size of 5 is selected in HTPDM and is replicated in WARD-HTP. A block size of 5 corresponds approximately to a real-world block dimension of 500km by 500km. The block dimension size is approximate because the mapping between coordinate degrees and kilometers is imprecise, and is dependent on location in the case of longtitudal points [13]. The approximate size of 500km is similar to the average hurricane total diameter [14], permitting the coarse-grain expression of hurricane trajectories without undue loss in precision.

All aspects of hurricane trajectory prediction with data mining depend on region discretization. Frequent trajectory patterns and their resulting trajectory predictions are expressed in terms of the regions where a given hurricane is expected to appear. The process of region discretization scales all hurricane coordinate trajectories downard, allowing the regionalized coordinates to be treated as a simplified coordinate system in subsequent aspects of the prediction model.

C. Trajectory Matching

The final key aspect of the hurricane trajectory prediction problem concerns accurate hurricane trajectory pattern matching. Pattern matching is a concern for hurricane trajectory prediction in two primary areas. First, hurricane trajectory matching must be effectively employed when comparing a new hurricane trajectory against the table of hurricane association rules generated by the previously described variant of AprioriAll. Hurricane trajectory matching is also used when checking the correctness of a trajectory prediction. The consequent of the matched hurricane trajectory rule is compared with the actual hurricane terminal trajectory to determine prediction correctness.

1) Rule-Fit Matching: Pattern matching is used when evaluating the best-fit rule for hurricane prediction purposes. Similar to HTPDM, a new hurricane trajectory is passed into the hurricane prediction model proposed in this work, it is separated into two subtrajectories: the initial trajectory T_i and the terminal trajectory T_t. T_i is compared against the antecedent of every interesting association rule satisfying minconf for a match. There are

two dimensions that determine the best-fit rule for an incoming trajectory requiring prediction. These dimensions are matching length and rule confidence. The matching length of the initial trajectory T_i to a rule antecedent is the longest congruent subsequence of regionalized points within the rule antecedent within T_i. Note that both T_i and the rule antecedent sequences have been regionalized per the regionalization methodology outlined. Using the fitness function described in HTPDM [4], the best-fit rule match for the incoming initial trajectory T_i is expressed with the following formula:

$$Match = max(conf_r * (1 - e^{-matchLen_{t,r}})), r \in R$$

Where r is a member of the set of all interesting assocation rules R, t is the initial trajectory T_i, and $matchLen_{t,r}$ expresses the matching length between T_i and the antecedent sequence of r. This formula is meant to be maximized to determine the rule which most generously fits the incoming hurricane trajectory. Once a best fit rule is determined, the rule consequent is compared with the *actual* terminal hurricane trajectory to determine if the prediction is correct.

2) Prediction Correctness Matching: Prediction Correctness checking is also powered by hurricane trajectory pattern matching. The consequent of the best-fit rule generated by the rule-fit matching stage constitutes the predicted terminal sequence for the incoming hurricane trajectory. The best-fit consequent is then compared against the actual terminal trajectory T_t to determine the prediction correctness. In HTPDM and WARD-HTP, prediction correctness is considered a binary attribute: a prediction is either correct or incorrect. If the matched-rule consequent also matches the incoming hurricane terminal trajectory, the prediction is correct and the model's correctness ratio is increased accordingly.

The pattern matching for rule consequents has some differences compared to antecedent pattern matching. First, prediction correctness uses matching length with a different definition. Whereas matching length for rule antecedents is the longest identical congruent subsequence of the initial trajectory, matching length for the rule consequent is the longest congruent subsequence of points within a specific distance, beginning at the start of the terminal sequence. This methodology is also based

on HTPDM and is based on the following steps. In the algorithm given, $conseq_r$ corresponds to the matched rule's consequent sequence, and D corresponds to the minimum distance threshold necessary for prediction correctness. In HTPDM, the distance threshold corresponds to a Euclidean distance of 1.

Algorithm 2 Consequent Matching for Prediction Correctness

1: $l \leftarrow min(|conseq_r|, |T_t|)$
2: $Select\ first\ l\ points\ of\ conseq_r$
3: $Select\ first\ l\ points\ of\ T_t$
4: **if** $All\ corresponding\ points\ within\ D$ **then**
5:     $Prediction : CORRECT$
6: **end if**

The hurricane trajecory prediction model based on data mining as originally proposed in [4] implements all aspects of the problem domain described. Modified AprioriAll, downscaled region discretization and pattern matching with respect to association rule antecedents and consequents are respectively employed by HTPDM in order to develop an effective system for predicting future hurricane trajectories. Based on Atlantic hurricane trajectory data from 1900 to 2008, the model generates a best-case correctness ratio of 65.0% and a worst-case ratio of 57.5%. The training data used in the original model consists of trajectories from 1900 to 2000, with testing data supplied by trajectories from 2001 to 2008. At a high level, the model is developed and tested by following the steps below, using all implementation strategies discussed.

1) Generate interesting trajectory rules from AprioriAll
2) Determine matching rule for incoming trajectory
3) Check if the matched rule's consequent points are sufficiently close

The differentiation between the model's best and worst-case hit rates is dependent on the ability of an incoming test trajectory to be matched with a trajectory association rule. As described in the pattern matching stage, the initial subtrajectory of an incoming trajectory is compared against all interesting rules to determine a best-fit match. If no such match is found, HTPDM attempts to extrapolate a

future point of the initial trajectory, using the extrapolated point as a new option for determing a rule antecedent match. When accounting for all tested trajectories, including those which were matched to a rule as a result of trajectory extrapolation, the correctness ratio was 57.5%. Therefore, the best case correctness ratio of HTPDM is the ratio which ignores all test trajectories which did not match with any rules, resulting in a final value of 65.0%.

Despite establishing a convincing initial start point, the best-case correctness ratio of 65.0% may be too low for HTPDM to be considered realiable in the real-world. This is problematic for those interested in hurricane trajectory prediction as a model with only 65.0% accuracy may not be able to deliver crucial insight into hurricane trajectory prediction, particularly with respect to landfall location prediction. Both correctness ratios given must be increased in order for data mining to be considered a practical approach to predicting the trajectories of future hurricanes.

V. SOLUTION

In order to increase the correctness ratio of datamining-based hurricane trajectory prediction, this work identifies and explores three opportunities for worst-case and best-case ratio refinement. The opportunities identified are the introduction of more realism in calculation formulas, reduction of historical impact on data used, and favorability towards recent hurricane trends. The opportunities for improvement are realized by three unique dimensions. First, the Haversine Distance Formula is employed for calculating realistic distances in spherical coordinate systems. Second, the use of weighted training data will be employed in AprioriAll to favor more recently recored hurricane trajectories. Finally, the training data used will only constitute hurricanes from 1950 onwards to remove the risk of relying on misrecorded historical data. The implementation of these three dimensions results in a new hurricane trajectory prediction model dubbed WARD-HTP (Weighted-Asset, Realistic-Distance Hurricane Trajectory Predictor). The combined contributions of WARD-HTP are expected to result in the desired increase to the best-case and worst-case correctness ratios.

A. Haversine Distance Formula

The first dimension of improvement introduced by WARD-HTP is the introduction of a realistic distance formula for hurricane trajectory comparison. Two different distance measurement formulas are explored in this work for use in the correctness checking stage of the hurricane trajectory prediction model. This work compares the Euclidean distance formula against the Haversine distance formula. The Euclidean distance formula (as used in [4]) is designed to determine the unit distance between two N dimensional points on a plane. In HTPDM, units correspond to degrees of latitude and longitude. The Euclidean distance formula is evaluated with two dimensions, corresponding respectively to latitude and longitude values.

In contrast to the Euclidean formula, the Haversine Distance Formula is designed to determine the kilometers between two latitude-longitude points over a large sphere, such as the earth's surface. Despite the fact that the earth's surface is not a perfect sphere [13], the Haversine Distance Formula is sufficiently accurate when used within a specific region such as the Atlantic Basin. The two formulas, denoted D_{eucl} and D_{havr} are defined as follows:

(1) $D_{eucl}(p, q) = \sqrt{(\Delta Lat)^2 + (\Delta Lon)^2}$

(2) $D_{havr}(p, q) = 2R(\text{atan2}(\sqrt{\alpha}, \sqrt{1 - \alpha}))$

$\alpha = \sin^2(\frac{\Delta Lat}{2}) + \cos(p_{lat}) + \cos(q_{lat}) + \sin^2(\frac{\Delta Lon}{2})$

Where ΔLat and ΔLon respectively denote the differences in latitude and longtitude between points q and p. Note that the Euclidean distance formula expresses Latitude and Longtitude in degrees and minutes, while the Haversine formula expresses Latitude and Longtitude exclusively in radians. In addition, whereas the Euclidean formula expresses distance in terms of units on a plane, the Haversine formula expresses distance in terms of portions of a 1-unit sphere. Therefore, to express distance in kilometers, the result of Haversine formula must be multiplied by R, which denotes the radius of the earth in kilometers (approximately 6371km [15]).

The euclidean distance in the discretized coordinate system corresponds to a difference in square regions approximately equal to the specified region dimension. Similar to the example outlined in

section 4.2, suppose the region dimension of the discretized coordinate system is once again 5. After discretizing the coordinate system into regions of square size, a Euclidean distance of 1 between two 5-degree square regions R_1 and R_2 means that R_1 and R_2 are approximately 1 region, or 5 degrees, or 500km apart.

When using the Euclidean formula, the distance threshold is 1. This corresponds to the distance threshold proposed in [4]. When the Haversine formula is employed, the maximal distance is varied from 300 to 450 km. This value corresponds to slightly less than approximately one average hurricane-diameter. Since the average hurricane diameter is between 333 and 670 km [14], two points are considered to be within the distance threshold if they are less than one average hurricane-diameter apart.

The maximal distance ranges used for the Haversine formula are much less than used in the Euclidean formula. Therefore, the expected outcome of employing the Haversine formula is the production of finer-grained hurricane trajectory prediction results when compared to results generated by use of the Euclidean distance formula.

B. Weighted Training Data

The second approach to increasing trajectory prediction correctness deployed in WARD-HTP is the usage of weighted training data in AprioriAll. Applying weights to recorded trajectories in chronological order allows AprioriAll to favor recently produced trajectory patterns by increasing the support and confidence of such recently produced trajectories. While the weighting of training data results in some trajectories being counted with a support greater than 1, this effect is balanced by counting earlier recorded trajectories with a support less than 1. In WARD-HTP, we weigh hurricane trajectories on a scale increasing linearly, resulting in a tipping point T where records before T are underweighted and records after T are overweighted. Depending on the weighing approach used, the weights of all records does not necessarily need to be balanced.

The following algorithm applies further modifications to AprioriAll in order to weigh the stored hurricane records. In this algorithm, W identifies

the total weighing scale to be used in the modification. For example, if W=3, then the final recored hurricane trajectory in the hurricane database HDB will be counted with a support of 3. In addition, the first trajectory will be counted with a support of 3/|HDB|. The support modifier increases by 3/|HDB| for each entry in HDB.

Algorithm 3 Modified AprioriAll With Weighted Support

1: **for** $(k = 1; L_{k-1} \neq \emptyset; k + +)$ **do**
2:      $w \leftarrow \frac{W}{|HDB|}$
3:      $C_k \leftarrow Candidates\ (Inner\ Join\ or\ Singletons)$
4:      $Prune\ C_k\ of\ infrequent\ subsequences$
5:      **for** $sequence\ S \in HDB$ **do**
6:          $Increase\ c \in C_k\ support\ by\ w\ if\ c \in S$
7:          $w \leftarrow w + \frac{W}{|HDB|}$
8:      **end for**
9:      $L_k \leftarrow \{c \in C_k|\ sup(c) \geq minsup\}$
10: **end for**
11: $FrequentTrajectories \leftarrow All(L_k)$

The algorithm described assumes that the HDB is ordered chronologically. This ensures that recently produced hurricane trajectory records correspond to a higher weight. Iteration through the HDB increases the current weight w by the weighting scale W divided by the size of the HDB. Therefore the final record in the HDB is guaranteed to have a weighted support exactly equal to W. Alternatively, the support weight modifier could also change every year. This means that W would be divided by the number of years which the database records span, and the value of w would increase every time the records reach a new year. Though index-based weighting is shown above in algorithm (3), both weighting strategies will be evaluated further in this work.

Under the assumption of chronologically ordered data, introducing weighted support to AprioriAll will result in higher support for patterns which occur in recent years. Fewer repeated records will be needed to generate frequent sequences from recent data, and more repeated records will be needed to generate frequent sequence from older data. Therefore, the introduction of weighted support to AprioriAll is expected to further increase the hurricane

trajectory prediction correctness ratio for current and future trajectories being tested upon.

C. Training Data Year Interval

The final improvement dimension proposed in WARD-HTP is the reconsideration of 1900 to 2000 as the trajectory prediction training data interval. In its original form, HTPDM is trained on hurricane trajectory data from 1900 to 2000. However, incorporating training data from years as early as 1900 introduces a risk of skewing frequent trajectory patterns in favor of historical errors and anomalies. While weighted support in AprioriAll helps reduce the impact of historical anomalies, the primitivity of early hurricane recording technology makes hurricane trajectories recorded before 1950 much less reliable than those recorded after [16]

The hurricane trajectory prediction model proposed in this work relies on test data exclusively from trajectories recorded between the years 1950 and 2000. The motivation being selecting this specific year interval is the attempted reduction of historical factors in the generation of frequent trajectory patterns (and subsequent interesting trajectory association rules). Paired with the Haversine Distance Formula and weighted AprioriAll approaches, tweaking the training data year interval is expected to result in an additional increase of the model correctness ratio. Additionally, association rules generated by training data from 1950 and onwards are expected to be more reliable as the influence of historical factors will be removed.

VI. EXPERIMENTAL SETUP

In its fully implemented form, WARD-HTP was run against various input parameters to determine an optimal configuration. WARD-HTP was evaluated against varying distance formula styles, weighted training data approaches (including the absence of weighted data), and training data year intervals. Evaluation of WARD-HTP was conducted on an eight-core Intel(R) i7-4720 processor at 3.35GHz, with 8 GB of memory using a Python implementation. The training and testing data was supplied from Hurricane DATabase 2 (HURDAT2), which is maintained by the National Oceanic and Atmospheric Administration [5].

Apart from the three dimensions of improvement outlined, one of the key differences between HT-PDM and WARD-HTP is that WARD-HTP does not attempt to extrapolate points for hurricane trajecories with which no rules were successfully matched. Therefore, results shown from WARD-HTP capture the correctness ratio when considering all test trajectories, and the correctness ratio when exclusively considering trajectory with which a rule was immediately successfully matched. Rule-match extrapolation was omitted from WARD-HTP due to the nature of the trajectories recored in HURDAT2.
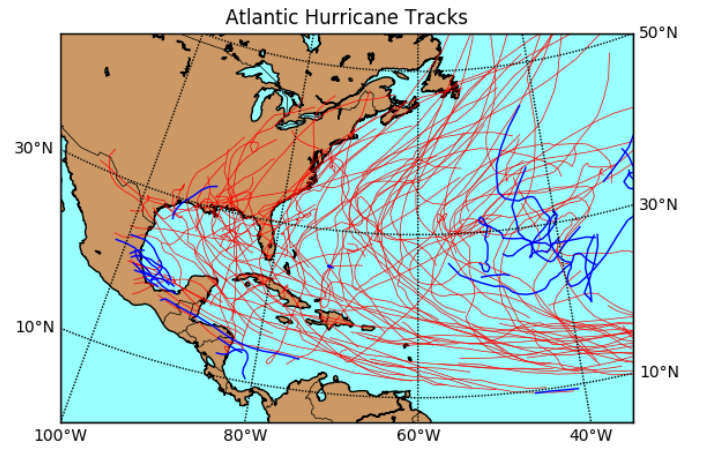
Fig. 3. Non-Matchable Hurricane Trajectory Paths

Figure (3) illustrates the characteristics of trajectories which were not able to match with any interesting rules. The particular diagram was generated from training data ranging from 1950 to 2000, and testing data from 2001 to 2015. Red trajectories represent a subset of the training data, while blue trajectories are those which did not match with any association rules. It is visually clear to see that non-matching trajectories were highly dissimilar from training data trajectories. As a result, no amount of point extrapolation would be able to result in a reliable terminal trajectory prediction. Some of the non-matching trajectories appear to line up with training trajectories, but are moving in an opposite direction. Therefore, any points they match with would also not provide a reliable result.

VII. EXPERIMENTAL RESULTS

This section of the paper documents the experimental results of all improvement dimensions proposed in WARD-HTP. Each dimension of Haversine Distance, Weighted Training Data and Training Data Year Interval of 1950-2000 contributes iteratively to a total growth in correctness ratio of 10.0% when compared to HTPDM, for a total of 75.0%. Whereas the worst-case correctness ratio (ratio accounting for rule-miss incidents) of HTPDM was 57.5%, the worst-case correctness ratio of WARD-HTP was 72.58% when configured in the same manner that generated the best-case result of 75.0%.

The worst-case to best-case difference in HTPDM is 7.5%, whereas the difference between the two cases in WARD-HTP is 2.44%. This results in a difference decrease of 5.08%. Not only is the difference in ratios reduced when using WARD-HTP, the best and worse case ratios are also increased, by 10.0% and 15.08% respectively.

A. Haversine Distance Formula

The Distance Formula section of the experiments in this work compare the Euclidean Distance approach as seen in HTPDM against two separate Haversine Distance approaches. The first approach imposes a distance threshold D of 300km while the second sets D to 450km. 300km was selected as a sample value for testing the Haversine formula, while 450km was selected as a value slightly less than the approximate average hurricane diameter discussed in section 4-B.

Aside from differences in the distance approaches, all other variables in this experiment are held constant. The model was trained on trajectories recorded between 1950 and 2000 and tested on tracjectories recorded between 2001 and 2015. The parameters of AprioriAll were set to minsup=3.75% (28.3125) and minconf=0.25. The training data in AprioriAll was unweighted (weighing scale=1), and rule-miss extrapolation was disabled as previous discussed.

As shown in figure (4), the Euclidean distance option resulted in much lower correctness ratios than both the 300km and 450km Haversine approaches. However, the difference between best and worst case correctness ratios is lowest when using Euclidean distance. Since increasing the ratios is prioritized over reducing the ratio difference, the Haversine approach of 450km produces the best result. The distance threshold of 450km introduces more flexibility into which predictions count as correct. The increase in flexibility is permissible as 450km still corresponds to a size small than the approximate average hurricane diameter. The results of this section show that using the Haversine Distance formula with a threshold of 450km is the ideal approach to evaluating distance when predicting hurricane trajectories with data mining.

B. Weighted Training Data

In this section of the evaluation against WARD-HTP, multiple full executions of the model from AprioriAll to the correctness ratio analysis are completed. Each execution of WARD-HTP is performed on a unique pairing of minimum support and the training data weighing strategy with the weighing scale W set to 1 or 2.5. Apart from the minimum support and the weighing strategy, all other parameters of the model are held constant.

Three different weighing strategies are evaluated in this experiment. First, the absense of data weighing is explored (W=1). Next, weighing on an annual scale is considered. In this strategy, the value of w increases by W divided by the total number of years in the data Y (in the case where training data is supplied from 1950 to 2000, Y=50). The third strategy employed is the index-weight strategy,

Distance Formula	Distance Threshold	Best-Case Ratio	Worst-Case Ratio
Euclidean	1	36.28%	33.00%
Haversine	300	56.19%	51.20%
Haversine	450	74.77%	68.10%

Fig. 4. Results of Different Distance Formulas

where the weight is updated at each HDB entry as seen in section 5-B.

Various values of minimum support are also explored in this series of evaluations. Minimum support values correspondong to 2.5%, 3.75% and 5% of the number of training data trajectories are paired against the three weighing strategies discussed above. While a higher minimum support will generate fewer interesting trajectory association rules, the rules generated will be more reliable for future predictions.

In figures (5) and (6), the number of frequent trajectories and trajectory association rules generated is compared alongride the best and worse care hit rates produced for given pairings of minsup and the weighing strategy. Note that the minconf value is held at 0.25 for all experiments, along with the training data of 1950-2000 from HURDAT2, testing data from 2001-2015, and distance formula set to Haversine Distance with a threshold of 450.

The maximal best-case hit rate for WARD-HTP was generated with an index-weighted strategy with minimum support of 2.5%, corresponding to 18.875 when used with data from 1950 to 2000. The resulting hit rate was 75.0%, with a worst-case of 72.58%.

Note that minimum support of 5% with index-based weighting generated a very slightly higher best-case hit rate at 75.22%, albeit with a lower worst-case ratio of 67.33%. We select 75% as the optimal generated data point because a .22% decrease in best-case hit rate is certainly a "fair trade" for a 5.25% increase in worst-case hit rate.

The year-based and index-based weighting strategies generated identical model results at minsup=2.5%. We select the year-based weighting strategy as the optimal approach due to the fact that the year-based strategy evenly counts trajectory records with respect to time. In the year-based strategy, all records from the most recent year are counted with at the maximum weight scale, whereas the index-based strategy only counts one record at the maximum. While the two weighting strategies resulted in identical results in this scenario, the year-based strategy is preferred as it encourages an increase in support weight closely coupled to real-world historical trends.

C. Training Data Year Interval

The final component of the evaluation undertaken in this work compares the correctness ratios generated by various training and testing data intervals.

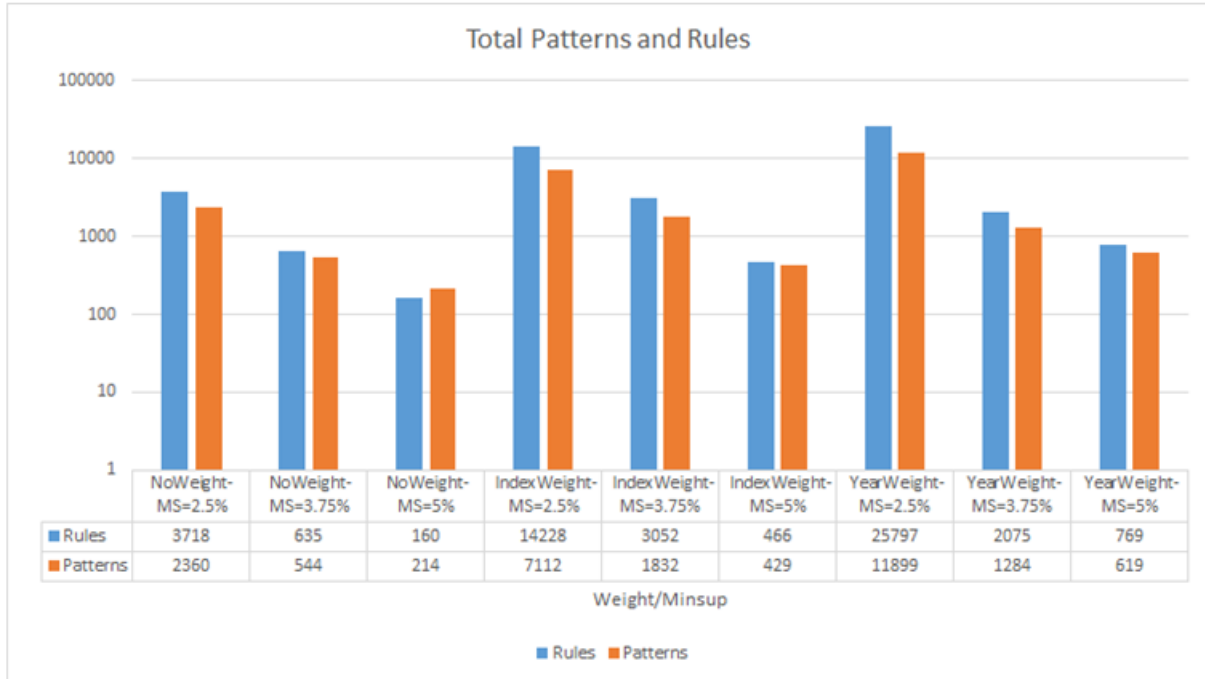|	NoWeight-MS=2.5%	NoWeight-MS=3.75%	NoWeight-MS=5%	IndexWeight-MS=2.5%	IndexWeight-MS=3.75%	IndexWeight-MS=5%	YearWeight-MS=2.5%	YearWeight-MS=3.75%	YearWeight-MS=5%
Rules	3718	635	160	14228	3052	466	25797	2075	769
Patterns	2360	544	214	7112	1832	429	11899	1284	619

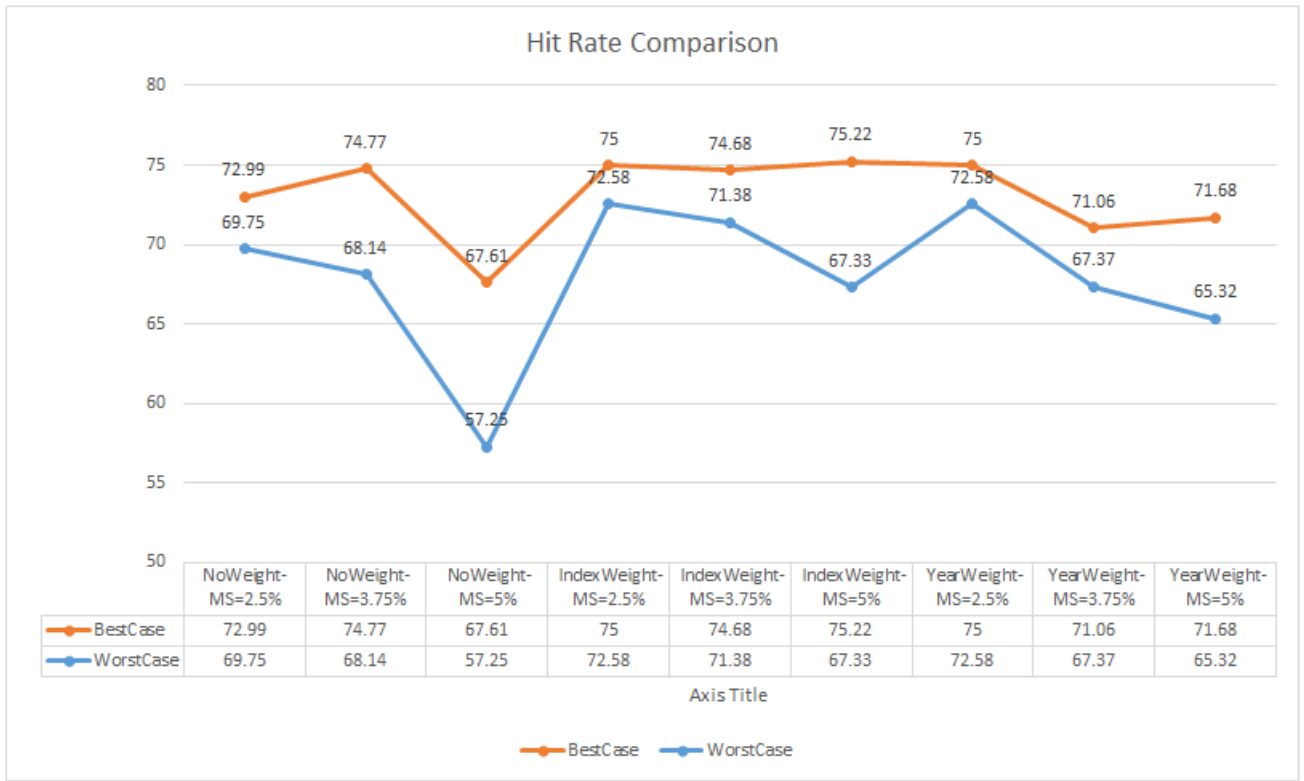Fig. 5. Minsup and Weighing Strategies Against Patterns and Rules

Fig. 6. Minsup and Weighing Strategies Against Correctness Ratios

Training data from 1900 to 2000 and 1950 to 2000 is used against testing data from 2001 to 2008 and 2001 to 2015. All other variables are held constant with minsup=2.5%, weighting strategy set to yearly and the distance formula set to Haversine with D=450. Note that the data ranges used by HTPDM were 1900-2000 for training data and 2001-2008 for test data.

Range	1900-2000	1950-2000
2001-2015	74.16 (71.77)	75.0 (72.58)
2001-2008	73.18 (70.62)	72.46 (69.93)

Fig. 7. Year Interval Options Against Correctness Ratios

Figure (7) shows the correctness ratios generated by all possibly training data and testing data pairings. Note that ratios in brackets denote the worse-case ratio while ratios outside of brackets refer to the best-case. Between all possible options, we find that the training data interval of 1950-2000 results in the best correctness ratios for testing data from 2001-2015 when applying all other parameters we

deemed optimal throughout the previous results of this work. Regardless of correctness ratio, 2001-2015 should always be selected as the test data interval to encourage analysis against more recent hurricane patterns. The results show that no training data-testing data pairing was able to generate correctness ratios higher than the ones generated by training the model with data from 1950 to 2000, and testing it on data as recently recorded as 2015.

VIII. CONCLUSIONS AND FUTURE WORK

This work proposed a new model named the Weighted-Asset, Realistic-Distance Hurricane Trajectory Predictor (WARD-HTP) for improving the accuracy of existing data mining approaches to hurricane trajectory prediction. An AprioriAll-based hurricane trajectory prediction model similar to [4] was developed to demonstrate the unique contributions of WARD-HTP. There are three such contributions brought forth in the development and implementation of the model proposed. These contributions are the use of training data from 1950 to 2000, the use of weighted training data, or

weighted assets (WA), and the use of the Haversine Distance Formula for realistic distance (RD). The three contributions combined form WARD-HTP, and produced a hurricane trajectory prediction correctness ratio of exactly 75.0%, which is 10% higher than the current state-of-the-art. When account for trajectories which were not able to match with any association rules, the correctness ratio of WARD-HTP is 72.58%, which is still 5.08% higher than the worst case state-of-the-art ratio of 57.5% as seen in [4]

There are mutliple areas to explore in the hurricane trajectory prediction problem space with respect to future work. Experiments in this work show that all datasets contain a subset of trajectories for which no association rule could be matched. While [4] proposed an extrapolation method to recover from rule-miss-incidents, more work must be completed to reduce the number of non-matchable trajectories. Many of the mismatched trajectories are outliers with very little conformance to existing hurricane trajectories as seen in figure (3). Additionally, more experiments will be run on different inputs including training data from other regions such as the Pacific region. Further approaches to increasing the correctness ratio of WARD-HTP will also be investiaged. This includes the investigation of alternative methods for defining training data weight.

REFERENCES

[1] Pielke Jr, Roger A., et al. Normalized hurricane damage in the United States: 19002005. Natural Hazards Review 9.1 (2008): 29-42

[2] F. F. Townsend, The Federal Response to Hurricane Katrina: Lessons Learned, United States Federal Government, Washington, DC, USA, 2006

[3] S. Sumathi and S. N. Sivanandam, Introduction to Dataming and its Applications: Springer Publishing 2006, ISBN: 978-3-540-34350-9

[4] Dong, X and C. Pi, D. (2013). Novel method for hurricane trajectory prediction based on data mining. Natural Hazards and Earth System Sciences. 13. 3211-3220. 10.5194/nhess-13-3211-2013.

[5] C. Landsea, J. Franklin and J. Beven (2015) The revised Atlantic hurricane database (HURDAT2). National Oceanic and Atmospheric Administration

[6] R. Agrawal and R. Srikant, Mining sequential patterns, Proceedings of the Eleventh International Conference on Data Engineering, Taipei, 1995, pp. 3-14. doi: 10.1109/ICDE.1995.380415

[7] Y. Su, S. Chelluboina, M. Hahsler and M. H. Dunham, A New Data Mining Model for Hurricane Intensity Prediction, 2010 IEEE International Conference on Data Mining Workshops, Sydney, NSW, 2010, pp. 98-105 doi: 10.1109/ICDMW.2010.158

[8] R. Yang, J. Tang and D. Sun (2011), Association Rule Mining Application for Atlantic Tropical Cyclone Intensity Changes, American Meteorological Society. doi: 10.1175/WAF-D-10-05029.1

[9] Rozanova, Olga S., Jui-Ling Yu, and Chin-Kun Hu. Typhoon eye trajectory based on a mathematical model: Comparing with observational data. Nonlinear Analysis: Real World Applications 11.3 (2010): 1847-1861.

[10] Chandan Roy and Rita Kovordanyi, Tropical cyclone track forecasting techniques: A review, 2012, Atmospheric research, (104-105), 40-69. doi: 10.1016/j.atmosres.2011.09.012.

[11] Kim, Hyeong-Seog, et al. Pattern classification of typhoon tracks using the fuzzy c-means clustering method. Journal of Climate 24.2 (2011): 488-508.

[12] J. D. Hunter, Matplotlib: A 2D Graphics Environment, in Computing in Science and Engineering, vol. 9, no. 3, pp. 90-95, May-June 2007. doi: 10.1109/MCSE.2007.55

[13] Menno-Jan, K. (2013). Map use: reading, analysis, interpretation: by Kimberling, AJ, Buckley, AR, Muehrcke, PC and Muehrcke, JO, Redlands, California, ESRI Press Academic, 2012. ISBN 978-1-589-48279-1 Seventh edition.

[14] D. Brown. (2013) Forecast Verification: Quantifying Forecast Uncertainty. National Hurricane Conference: Hurricane Readiness for Coastal Communities

[15] Chambat, F., and Valette, B. (2001). Mean radius, mass, and inertia for reference Earth models. Physics of the Earth and Planetary Interiors, 124(3), 237-253.

[16] Landsea, C. W., Vecchi, G. A., Bengtsson, L., and Knutson, T. R. (2010). Impact of duration thresholds on Atlantic tropical cyclone counts. Journal of Climate, 23(10), 2508-2519.