Data Mining Historical Weather Patterns at Valentia Observatory (1992-2021)

BARRY, Patrick

Abstract

Analysis of historical meteorological data is crucial for understanding climate patterns. This project applies data mining techniques to nearly 30 years (1992-2021) of hourly weather data from Valentia Observatory, Ireland. Challenges including addressing a significant sensor change in 2012. K-Prototypes clustering and Apriori association rule mining were employed on the pre-2012 data using R. The analysis successfully identified five distinct and meteorologically interpretable weather profiles and quantified significant co-occurrence patterns through association rules, such as strong seasonal temperature links. This work demonstrates the utility of data mining in uncovering complex, meaningful structures within long-term weather datasets beyond simple statistical summaries.

1. Introduction

1.1. Context

Analyzing historical weather data is crucial for understanding local climate patterns and broader environmental changes. This project focuses on data from a weather station in rural Ireland to find unexpected patterns from the past 30 years.

Valentia Observatory, the weather station used for this project, made its first weather observation in 1860 (1). This project uses the data recorded over almost 30 years from 1992 to 2022.

On the first of April 2012, the weather station switched from manual to automatic observations. Due to this, several variables stopped recording data on this date, necessitating careful handling during analysis.

1.2. Project Goal

The goal of this project is to apply data mining techniques such as clustering and association rule mining using R to uncover trends, patterns, and relationships among the various variables available in the dataset.

1.3. Motivation

I chose this dataset to gain insight into the weather patterns near my homeplace, and try to uncover explanations for the local climate. It is also a good challenge to handle time-series data, and a practical real-world problem.

2. Data Understanding

2.1. Data Source

This data, originally collected by Met Éireann (the Irish Meteorological Service), was obtained from Kaggle.

The first observation is from the first of January 1992 at 00:00:00, and the last observation is from the first of February 2022 at 00:00:00. The observations are made hourly, meaning there is 263,736 observations in the dataset.

2.2. Data Overview

Of the 21 variables, key variables include measurements of temperature (air, wet bulb, dew point), precipitation, sea level pressure, humidity, wind speed and direction, sunshine duration, visibility, and cloud properties.

”Wet Bulb Air Temperature” is the lowest temperature air can reach by evaporating water, is an indicator of humidity and heat stress. ”Dew Point Air Temperature” is the temperature at which air becomes saturated with moisture, leading to condensation. ”Vapour Pressure” is the pressure exerted by water vapor in the air, showing moisture content. ”Relative Humidity” is the percentage of moisture in the air compared to its maximum capacity at a given temperature.

2.3. Data Quality

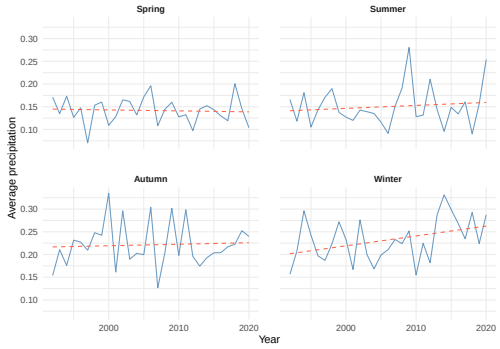
Investigating the data, I realised several variables had missing data after the switch to an automatic weather station, namely: present and past weather, sunshine amount, visibility, cloud height, cloud amount. I decided to split the dataset into two subsets, one with all the data from the 30 years but none of the affected variables, and another one with all variables up to April 2nd 2012. Different analysis will be performed on these two datasets during the project.

The core variables of temperature, precipitation, wind speed,

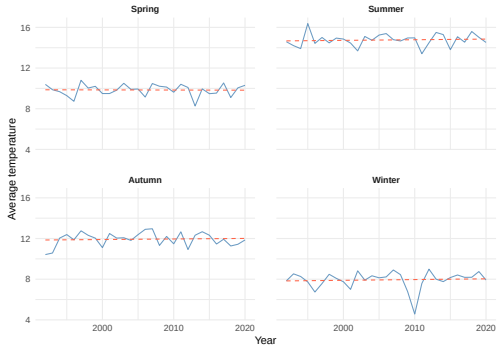
humidity, and more all had less than 0.0004% of their values missing. The process to handle this is detailed further in the Data Preparation section below.

Indicator variables are also present in the data, which will require decoding to be used in the project. These variables tell the confidence with which a measurement is made (e.g. rainfall indicator). The present and past weather variables will have to be decoded also, more detail on that in the Data Preparation section.

2.4. Data Exploration



(a) Precipitation



(b) Temperature

Figure 1. Yearly trends by season for different weather variables. Data covers full seasons up to Autumn 2021.

Figure 1a displays the yearly average precipitation trends within each season. Significant year-to-year fluctuations are evident across all seasons. The linear trend lines suggest a noticeable increase in average precipitation during Summer and Winter over the period analysed. Autumn also shows a slight increasing trend, while the trend in Spring appears to be a slight decreasing.

Figure 1b illustrates the temperature change over the past 30 years. The values are much more stable than the precipitation values, yet we still see one exceptionally cold year in Winter 2010, which was in fact the coldest winter on record for Ireland, with heavy snowfall (2). On average, the increase in temperature is negligible per season, especially compared to precipitaion amounts.

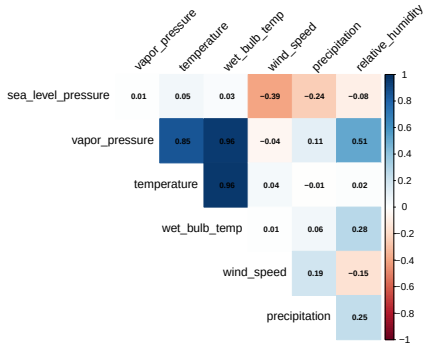


Figure 2. Correlation Matrix of Core Weather Variables

The correlation matrix in Figure 2 shows the relationships between key weather measurements. Very strong positive correlations (r = 0.96) can be seen between temperature and both wet bulb temperature and vapor pressure, indicating these variables capture closely related aspects of air temperature and its moisture-holding capacity. Vapor pressure also shows a moderate positive relationship with relative humidity (0.51). On the other hand, sea level pressure displays a moderate negative correlation with wind speed (-0.39), consistent with lower pressure often accompanying increased wind. Most other variable pairs only show weak correlations.

3. Data Preparation

Based on the data understanding phase, several preparation steps were performed using R, primarily using the `dplyr` and `lubridate` packages, to clean the data and prepare it for analysis:

- **Column Renaming and Type Conversion:** First steps involved renaming the original columns to more descriptive names (e.g., `rhum` to `relative_humidity`) for clarity. The `date` column was parsed from its character format into a standard date-time object using `lubridate` to allow temporal analysis.

- **Dataset Splitting:** A crucial observation was the switch from manual to automatic observations on April 2nd, 2012. This resulted in several variables (specifically `present_weather`, `past_weather`, `sunshine_duration`, `visibility`, `cloud_height`, and `cloud_amount`) having no data recorded after this date. To manage this, the dataset was split into two subsets:

weather_data_pre_2012 containing all variables but only data up to April 1st, 2012, and weather_data_complete, excluding the sensor-affected variables.

- **Handling Missing Values and Indicators:**
  - Indicator variables (irain, itemp, iwb, iwdsp, iwddir) sometimes contained undocumented or missing codes (e.g., irain = −1). These were handled by either converting to NA, or imputed based on related present weather codes.
  - As stated in the Data Quality section, the core variables of temperature, precipitation, wind speed, humidity, and more all had less than 0.0004% of their values missing. These missing values were imputed based on hourly and monthly means or modes specific to each variable.
  - The numeric indicator variables themselves were decoded from their integer codes into meaningful factor levels representing the observation quality or type (e.g., irain = 0 became "Satisfactory", irain = 6 became "Estimate Trace Precip").

- **Decoding Weather Codes:** The numeric SYNOP codes in present_weather (ww) and past_weather (w), ranging from 0-99, were decoded. For present_weather, this created two new factor variables: present_weather_category (grouping similar conditions like 'Rain', 'Fog', 'Drizzle') and present_weather_intensity ('Light', 'Moderate', 'Heavy', 'Extreme'). For past_weather, only the past_weather_category was created.

- **Cloud Height Transformation:** The special code 999 in the cloud_height variable, signifying no cloud ceiling observed, was converted to NA to represent missing data in the standard way. A binary variable, has_cloud_ceiling, was also created (1 if cloud_height was not 999, 0 otherwise).

- **Feature Engineering:** New temporal features were extracted from the parsed date column to analyse cyclical patterns. These included: hour, day_of_week, month, year, week_of_year, and season.

- **Temporal Subsetting:** Finally, to ensure consistent yearly comparisons in trend analysis and aggregations, the partial data from January 2022 was removed from the primary analysis datasets (weather_data_complete and weather_data_pre_2012).

4. Modeling

In this section, several data mining techniques were applied to the weather dataset. K-Prototypes clustering is an unsupervised method which can be used with both numerical and categorical features to identify distinct groups or clusters within the data. Association Rule Mining with the Apriori algorithm was implemented to discover relationships between different weather variables.

4.1. Clustering (k-Prototypes)

To identify distinct types of hourly weather conditions present in the pre-2012 dataset, clustering was performed using the k-Prototypes algorithm (3). This unsupervised method is well-suited for grouping observations based on similarity considering the dataset's mix of numerical and categorical features.

The clustering was performed on a selection of key variables. Numerical variables included: temperature, precipitation, relative_humidity, wind_speed, sea_level_pressure, and sunshine_duration. Categorical variables included: present_weather_category, present_weather_intensity, has_cloud_ceiling, and month. These variables were chosen as the key factors influencing weather patterns, representing temperature, precipitation, wind, and seasonal timing.

The optimal number of clusters (k) was determined using the Elbow method. This involved running the k-Prototypes algorithm for k values from 2 to 8 and plotting the Total Within-Cluster Sum of Squares (WCSS) against k. The 'elbow' point on this plot, where the rate of decrease in WCSS decreases, suggests an appropriate balance between cluster cohesion and model complexity. Figure 3 shows the elbow curve for this analysis. Based on the plot, k=5 was selected as the elbow point, or the optimal number of clusters.

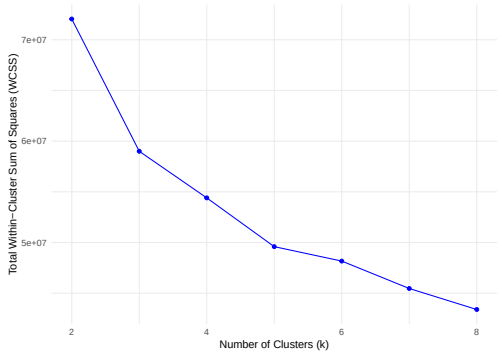


Figure 3. Elbow Method for Determining Optimal k (k-Prototypes).

The final k-Prototypes model was then executed with the chosen k=5. The `lambda` parameter, which balances the contribution of numerical and categorical variables, was estimated from the data using the `lambdaest` function within the `clustMixType` package (4). To improve the quality of the clustering solution, the algorithm was run with `nstart=5` (meaning it initiated from 5 different random starting points and retained the best result) and allowed a maximum of `iter.max=50` iterations for convergence within each start. The execution time for fitting the final k=5 model was around 4 minutes on my laptop. The resulting cluster assignments for each data point were stored for evaluation.

4.2. Association Rule Mining (Apriori)

To discover relationships and patterns of co-occurrence among weather conditions within the pre 2012 dataset, Association Rule Mining (ARM) was employed using the Apriori algorithm (5), implemented with the `arules` R package (6). The objective was to find rules in the form "If condition X occurs, then condition Y is likely to occur."

A different subset of variables was selected for this algorithm. Categorical variables included `has_cloud_ceiling`, `month`, and `hour`. Numerical variables considered were `temperature`, `precipitation`, `relative_humidity`, `wind_speed`, and `sea_level_pressure`.

The Apriori algorithm requires itemsets (categorical data), so the following preprocessing was necessary:

- **Discretization:** The selected numerical variables (`temperature`, `precipitation`, etc.) were converted into categorical bins. This was done using the `discretize` function with `method = "frequency"` and `breaks = 3`, creating three bins of approximately equal size for each variable, labeled as 'Low', 'Medium', and 'High' (e.g., `temperature_Low`, `temperature_Medium`, `temperature_High`). This transforms continuous measurements into distinct items suitable for rule mining.
- **Transaction Format:** The data frame of only categorical or discretized variables, was then converted into the `transactions` data structure using the `as(..., "transactions")` method, which is the required input format for the `arules` package functions.

The Apriori algorithm was then executed on the transaction dataset with the following parameters:

- `support = 0.05`: This minimum support threshold requires that the itemset (combination of conditions in a rule) must appear in at least 5% of all hourly observations (transactions).
- `confidence = 0.6`: This minimum confidence threshold means that for a given rule $\{\text{LHS}\} \Rightarrow \{\text{RHS}\}$, at least 60% of the transactions containing the Left-Hand Side (LHS) conditions must also contain the Right-Hand Side (RHS) condition.
- `minlen = 2`: This ensures that only rules involving at least two items (conditions) are generated.

The algorithm generated a total of 96 association rules satisfying these criteria. The computation time for running the Apriori algorithm was less than 0.01 seconds. These rules were then available for inspection and evaluation to potentially identify interesting weather patterns.

5. Evaluation

This section evaluates the patterns discovered through clustering and association rule mining, assessing if meaningful insights were discovered and if the methods produced good results.

5.1. Clustering Results

The k-Prototypes algorithm identified five clusters in the pre-2012 data. Cluster sizes were: C1: 48.2k (27.2%), C2: 36.3k (20.5%), C3: 24.7k (13.9%), C4: 36.1k (20.4%), C5: 31.8k (17.9%). Cluster 1 was most frequent, Cluster 3 least. The cluster centroids revealed distinct profiles:

- **Cluster 1 (Calm, Humid, Overcast - High Pressure):** Most common. Defined by lowest wind speed (5.2 kt), high pressure (1022.9 hPa), high humidity (87.5%), minimal sunshine (0.09 hrs), and low precipitation (0.01 mm). Modal in August. Represents settled, humid, overcast conditions.
- **Cluster 2 (Warm, Dry, Sunny - High Pressure):** Fair weather. Lowest humidity (66.9%), highest sunshine (0.37 hrs), negligible precipitation (0.006 mm), high pressure (1023.0 hPa), and moderate wind (9.0 kt). Modal in May. Represents warmer, drier, sunnier periods.
- **Cluster 3 (Cool, Wet, Windy - Low Pressure):** Least frequent 'bad weather'. Highest precipitation (0.73 mm/hr), highest wind (13.0 kt), lowest pressure (995.3 hPa), cool temperatures (10.2°C), high humidity (88.5%), and minimal sunshine (0.04 hrs). Modal weather 'Rain', modal month October. Represents typical Irish stormy conditions.

- **Cluster 4 (Cool, Breezy, Partly Sunny):** Moderate conditions. High wind (11.7 kt), moderate sunshine (0.20 hrs), moderate-low pressure (1006.0 hPa), low precipitation (0.03 mm), and moderate humidity (75.7%). Modal in April. Represents cool, breezy days with broken cloud.
- **Cluster 5 (Mild, Very Humid, Drizzly):** Warmest temperatures (12.3°C), highest humidity (93.3%), moderate precipitation (0.39 mm/hr, modal 'Drizzle'), lowest sunshine (0.02 hrs), moderate wind (9.8 kt) and pressure (1015.9 hPa). Modal in July. Represents mild, very damp, overcast conditions.

The clusters were visualized using t-SNE (Figure 4).

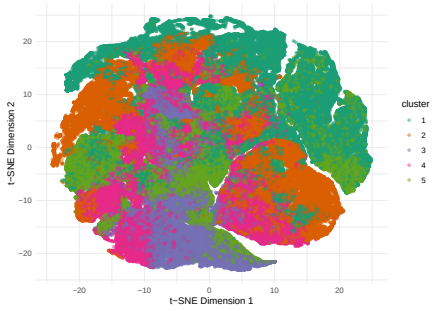


Figure 4. t-SNE Visualization of Weather Clusters (k=5).

The t-SNE plot shows there are distinct concentrations despite the expected overlap. Cluster 1 (Dark Green) appears to be the most separated, representing mild conditions. Cluster 2 (Orange) also shows some distinction for fair weather. Other clusters (Pink-4, Green-5) show more mixing, potentially representing more common states. The visualisation supports the presence of distinct underlying structures in the data.

Overall, the clustering successfully identified five interpretable weather profiles, providing a richer understanding of typical patterns at Valentia Island and achieving good results in segmenting the data.

5.2. Association Rule Results

The Apriori algorithm generated 96 rules (support greater than 5%, confidence greater than 60%), revealing common co-occurrences. Key patterns include:

- **Seasonal Temperatures:** Strong rules confirmed expected links, e.g., "July" ⇒ "High Temperature" (Conf=87.5%, Lift=2.62) and "January" ⇒ "Low Temperature" (Conf=69.1%, Lift=2.11).
- **Calm Conditions Pattern:** The rule "Low Temp., High Humidity" ⇒ "Low Wind" (Conf=69.2%, Lift=2.38) highlights that cold, damp conditions are 2.38 times more likely than average to also have low wind speeds, possibly indicating stable air or fog potential.

A network graph visualizes the strongest associations (Figure 5).

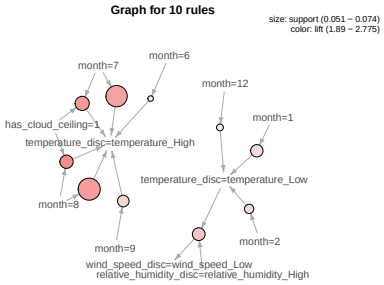


Figure 5. Network Graph of Top 10 Association Rules by Lift.

Figure 5 highlights the strong month-temperature links and the other key relationships discovered.

In summary, association rule mining verified expected seasonal patterns and revealed more specific conditional relationships, complementing the cluster analysis despite the loss of information by discretisation.

6. Deployment

The findings from this data mining project offer several potential applications. The identified cluster profiles provide a nuanced understanding of typical weather patterns at Valentia Observatory beyond simple averages, characterising distinct conditions like "Calm, Humid, Overcast" (Cluster 1), "Warm, Dry, Sunny" (Cluster 2), and "Cool, Wet, Windy" (Cluster 3). This improved understanding of the local microclimate could be valuable for regional planning, agriculture, or tourism. Furthermore, specific association rules, such as the one linking low temperature and high humidity to low wind speeds, could help identify conditions conducive to specific phenomena like fog formation, potentially informing local forecasting or advisories.

The R code used for data preparation, analysis (including clustering and association rule mining), and visualization, along with the processed datasets, are available on GitHub: `https://github.com/patbarry29/ireland-weather-data`. Key R packages utilised include `dplyr`, `lubridate`, `clustMixType`, `arules`, `arulesViz`, `ggplot2`, and `Rtsne`.

Potential future work could involve using the derived cluster labels as target classes for predictive modeling, attempting to forecast the likely weather type for upcoming hours.

Another avenue would be to compare trends in the core variables (available throughout the period) before and after the 2012 sensor automation to assess any systematic changes, or to investigate the temporal evolution in the frequency of the identified weather clusters over the three decades.

7. Conclusion

This project successfully applied data mining techniques to nearly 30 years of hourly weather data from Valentia Observatory. After extensive data preparation, which included handling a sensor change in 2012, decoding variables, and feature engineering, K-Prototypes clustering and Apriori association rule mining were performed on the pre-2012 dataset.

The analysis revealed meaningful patterns and relationships within the local weather system. Key findings include the identification of five distinct and meteorologically interpretable hourly weather profiles, such as typical high-pressure calm/overcast conditions, fair sunny weather, low-pressure stormy conditions, breezy partly sunny days, and mild drizzly periods. Association rule mining confirmed expected strong seasonal temperature links and uncovered more specific relationships, like the increased likelihood of low wind speeds during cold, damp conditions. We can conclude that data mining techniques effectively discovered insightful structures and dependencies embedded within this historical weather data.

The most valuable insights gained include the characterisation of these distinct weather clusters, providing a deeper knowledge of local weather than simple statistics allows, and the discovery of certain conditional patterns through association rules, such as the link between low temperature, high humidity, and low wind speed.

The primary limitation of this study was the 2012 sensor change, which restricted the analysis of several key variables (like sunshine, visibility, detailed weather codes) to the pre-2012 period. The requirement to discretise numerical variables for association rule mining also represents a simplification of possible continuous relationships. Despite these limitations, the project demonstrated the utility of data mining for extracting valuable knowledge from complex historical datasets.

References

[1] M. Éireann. (2025) Valentia observatory. [Online]. Available: https://www.met.ie/about-us/our-history/valentia-observatory

[2] ——. (2025) Extreme cold spell. [Online]. Available: https://www.met.ie/cms/assets/uploads/2017/08/ColdSpell10.pdf

[3] Z. Huang *et al.*, "Clustering large data sets with mixed numeric and categorical values," in *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)*. Citeseer, 1997, pp. 21–34.

[4] G. Szepannek, "clustmixtype: User-friendly clustering of mixed-type data in r," *The R Journal*, pp. 200–208, 2018. [Online]. Available: https://doi.org/10.32614/RJ-2018-048

[5] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th International Conference on Very Large Data Bases*, ser. VLDB '94. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, p. 487–499.

[6] M. Hahsler, S. Chelluboina, K. Hornik, and C. Buchta, "The arules r-package ecosystem: Analyzing interesting patterns from large transaction datasets," *Journal of Machine Learning Research*, vol. 12, pp. 1977–1981, 2011. [Online]. Available: https://jmlr.csail.mit.edu/papers/v12/hahsler11a.html