# Interim Report

**Jamie Elliott**
110215415
School of Computing
Newcastle University
1, Urban Sciences Building, Science Square, Newcastle upon Tyne NE4 5TG
t.j.elliott@newcastle.ac.uk

# 1 Introduction

The term "industrial coating" describes paint or paint-like material that is applied to a surface to provide enhanced properties to the surface. Valued at USD 80.87bn in 2019 and forecast to grow to USD 107.32bn by 2027 the industrial coatings industry is a huge market globally [globenewswire.com, 2020]. There are many end use industries for industrial coatings including marine, vehicle refinish, automotive, electronics, oil and gas, mining, aerospace, and renewable energy. The surfaces coated can range from minuscule electrical components to huge infrastructure such as bridges and stadiums. Industrial coatings are primarily used for their protective properties. The protective properties generally come from being resistant to corrosion, weathering and physical damage. Atmospheric Corrosion is the attack of a metal by the atmospheric environment to which it was exposed [de la Fuente et al., 2011] and is an important factor to considers when evaluating the protective properties of an industrial coating. Atmospheric corrosion is caused by rainwater or condensing water, oxygen and atmospheric pollutants [LeBozec et al., 2015] and therefore effects a multitude of coated surfaces. In order to better understand the resistance of a coating to the atmospheric environment coatings companies use "exposure testing sites" (ETS) where coated substrates are left exposed to the atmosphere for anywhere between weeks and years and then performance is assessed. AkzoNobel has many of these test sites around the world in a variety of different locations(Felling, Sunderland, Plymouth, Ruco Grande, Sanary, Florida, Houston, Phoenix, Dammam, Bangalore, Geoje, Suzhou, Pudong, Suzhouou, Songjiang, Willawong, Melbourne). The aim of these test sites it to be representative of the in service environments that AkzoNobel's industrial coatings are exposed to. Currently ETS and in-service environments are classified at a basic level e.g Cold/Dry/ High UV or Tropical/Wet/High UV but it is thought that this rudimentary classification is not detailed enough to provide accurate insight into product performance on a global scale. With the rise of readily available global climate data and the development of Data Science techniques to analyse this data an opportunity to carry out a more detailed investigation of the environments of the ETS and how well they represent global in-service environments has arisen. With a better understanding of how the atmospheric environment relates to product performance significant improvements can be made to product testing, product specification and R&D objectives. Product testing that is more representative of in-service environment on a case by case basis will be identifiable. Bespoke product specification to each customer depending on the location of the asset will be possible. Further in the future R&D objectives will be identifiable based on an improved understanding of product performance related to environment.

# 2 Aims and Objectives

## 2.1 Aim

The aim of the project is to understand the environments of AkzoNobel's ETS and how they relate to the range of environments across the globe.

## 2.2 Obejctives

### 2.2.1 Build relevant climate data set

The first step in achieving the aim of the project is to define which climatic variables are relevant to atmospheric corrosion. This definition will be informed by industry expertise from within AkzoNobel as well as a review of relevant academic literature. Once relevant data has been identified this will be prepossessed to an appropriate spatio-temporal resolution taking into account the level of detail required and also ensuring a manageable computational requirement. Climactic data on a global scale is typically provided in raster format. A raster is a matrix of cells arranged in rows and columns where each cell contains a value representing the information. These rasters can be layered on top of each other to create a multivariate, spatial data set.

### 2.2.2 Compare the climate of exposure test sites to each other

Hierarchical Clustering on Prinicipal Components will be carried to identify which ETS are similar to each other and at what point in the year. This will provide an opportunity to optimise future testing programs based on avoiding duplicating testing in two ETS which are significantly similar to each other.

### 2.2.3 Compare climate of exposure test sites to global climates

The climates of each ETS will be compared to global environments. Analysing the maximum similarity level of each cell in the raster when compared to all ETS will highlight global climates not covered by current ETS.

### 2.2.4 Identify the optimum number and location of exposure test sites

"Add one in" modelling to suggest a location for a new ETS that significantly improves the global climate coverage of AkzoNobel ETS according to this measure of similarity.

## 3 Overview of Progress

### 3.1 Build relevant climate data set

The first step in carrying out this analysis is to identify variable relevant to atmospheric corrosion of industrial coatings in order to create the working data set. The required data set was determined after consultation with industry experts along side a review of academic literature [Daly, 2006]. Key variables identified are Temperature, Precipitation and Concentration of electrolyte [LeBozec et al., 2018]. Atmospheric corrosion is an electrochemical process therefore, anodic and cathodic electorlytes must occur in the presence of water [de la Fuente et al., 2011]. Temperature typically accelerates electrochemical processes. Precipitation facilitates the reaction by providing water, however, a large amount of precipitation could slow down the reaction by diluting the concentration of the electrolytes. Finally, electrolytes must be present for the reaction to occur therefore a high concentration in the atmosphere is more likely to produce atmospheric corrosion [LeBozec et al., 2018]. In addition to these variables it is vital to take into account the time of wetness (TOW) and frequency of wetness (FOW) a surface experiences. A high TOW means that there is a large amount of time corrosion can occur. The frequency of wet and dry periods also effects the severity of corrosion. A high frequency of change is significantly more aggressive to a coating or substrate compared to extended periods of wet or dry before changing [LeBozec et al., 2018].

The Copernicus Climate data store (CDS) will be used as the primary data source for this project. CDS provides global Temperature, Precipitation and Sea salt aerosol concentration data from 1850-2300 and in a variety of spatial resolutions[ECMWF, 2016]. QGIS is an open-source application that is often used for geospatial data analysis and will be utilised in this project. The spatial resolution chosen will depend on computational resources available during the project and is still to be determined.

TOW is not a readily available data set. However, it is possible to estimate TOW using the method described in Schindelholz and Kelly [2012] that states "the total time when the relative humidity (RH) of the ambient environment is greater than or equal to 80 % at temperatures above 0 ° C.". Using this method TOW for all global locations will be calculated. FOW is also not readily available and there exists no standard method for determining this value. It is yet to be determined how FOW will be calculated for this analysis.

### 3.2 Compare the climate of Exposure testing sites to each other

The approach to be adopted for this objective has been identified through a literature review of climatic clustering. Similar approaches are identified in much of the literature relating to clustering of climates [Praene et al., 2019, Netzel and Stepinski, 2016, Kozjek et al., 2017, Park et al., 2019, Gönençgil, 2020]. Most studies of this nature follow a similar methodology, the climate of a location is represented by long-term monthly mean for a set of climactic variables, a dissimilarity function based on Euclidean distance is then fed into a clustering algorithm in order to generate clusters of similar climates. Often PCA or a similar factor analysis technique is used to reduce the dimensionality of the vectors representing each individual location prior to measuring dissimilarity and clustering. Praene et al. [2019] provides a title for this technique "Hierarchical Clustering on Principal Components".Detailed in Netzel and Stepinski [2016] is a discussion of the most effective method for global climate classification while making special considerations for climate data. In this paper the best methods for normalisation, dissimilarity function and clustering algorithm are discussed. In the case of normalisation it is vital to take into account the skewness of precipitation data. Precipitation data has a distribution that skews significantly towards high values meaning most values after normalisation would be zero and the influence precipitation data has on clustering would be reduced. In order to avoid this an alteration must be made to the normalisation function. The altered normalisation of precipitation values is outlined in Figure 1 where "R" is the precipitation value. All other normalisation will follow standard normalisation procedure Xi= [Xi - min(Xi)]/ [max(Xi) - min(Xi)] where Xi is the variable value.Often local climates are represented using long-term monthly means. However, in Netzel and Stepinski [2016] local climates are represented as cyclical time series. Netzel et al believe that this representation better accounts for the month to month sequence of climates and more closely represents human perception of weather. In addition, this represntation allows the utilisation of techniques used to analyse time series data that prove to be more effective for clustering climate data. In Netzel and Stepinski [2016] the most effective dissimilarity measure used is Dynamic Time Warping (DTW). DTW is a technique used for measuring similarity between time series which may vary in speed to find the optimal match for each data point. In the case of annual climate data this means that data can be compared to each other out of sequence. For example The most similar climate to Melbourne in January might be Florida in August. Traditional euclidean distance measurement would only compare climates for the ith point in each time series i.e Melbourne in January to Florida in January therefore this would not account for the similarity in climate stated above. Finally, the preferred clustering method identified is Partitioning around medioids (PAM). Taking into account the information provided in these research papers the analytical approach taken in this project will follow the Hierarchical Clustering on Principal Components outline by Praene but will incorporate the altered normalisation detailed in Netzel and will also make use of PAM as opposed to K-means to carry out clustering. Time allowing, a separate analysis utilising DTW as the Distance measure will be undertaken and the results of these two analysis will be compared. To gain an out put of similarity on a monthly basis using DTW time series of daily data for each month must be created, this significantly increases the computational requirements of pre-processing and analaysis and a solution based on cloud computing would be required. Should the two methods of analysis be comparable the euclidean distance measure will be selected as this is more efficient from a computational perspective.

### 3.3 Compare climate of Exposure testing sites to global climates

Following on from a comparison of ETS to each other each testing site will be compared to global environment. This will be carried out using the rasterised climate data comparing each test site to each cell in the raster data. The comparison will be based on the euclidean distance between a test site and each cell. This analysis will allow comparison of global environments for the most similar test site. In addition, comparing the maximum similarity value of each cell in the grid to any single testing site will identify "blind spot" climates that are not currently covered by ETS.

### 3.4 Identify the optimum number and location of exposure test sites

Following on from blind spot identification this project will conduct add one in modelling to identify potential locations of future ETS that would significantly improve the global climate coverage of product performance testing. Each potential location will be compared to each cells in the raster data set and the locations with the most improvement in similarity for climates not currently covered will be proposed as future ETS sites.

## 4 Evaluation

There are multiple methods for evaluating the quality of the clustering obtained. The most popular is Homogeneity and V-measure. Homogeneity(H) and Completeness(C) contribute to V-measure. Homogeneity is the objective that a cluster only contains members of a single class. Completeness is the objective that all members of a given class are assigned to the same cluster. V-measure is then defined as 2x(HxC/H+C) [Rosenberg and Hirschberg, 2007]. Another method of evaluating quality of clustering is ‘Explained predictand variance’ this is a measure of the within-cluster sum of squared distances. This minimises the impact of data points that are close together and maximises the input of data points that are further apart within a cluster[Zscheischler et al., 2012]. These methods will be used to evaluate the robustness of the statistical analysis. In order to evaluate the use of this method in industry future exposure testing of industrial coatings is necessary. Unfortunately due to the time scale of this project this testing will not be able to be carried out along side the statistical analysis.

# 5 Appendices

$$R \leftarrow \begin{cases} \frac{R}{350}, & \text{if} \quad R \leq 350 \\ 1, & \text{if} \quad R > 350 \end{cases}.$$

Figure 1: Normalisation for Precipitation data

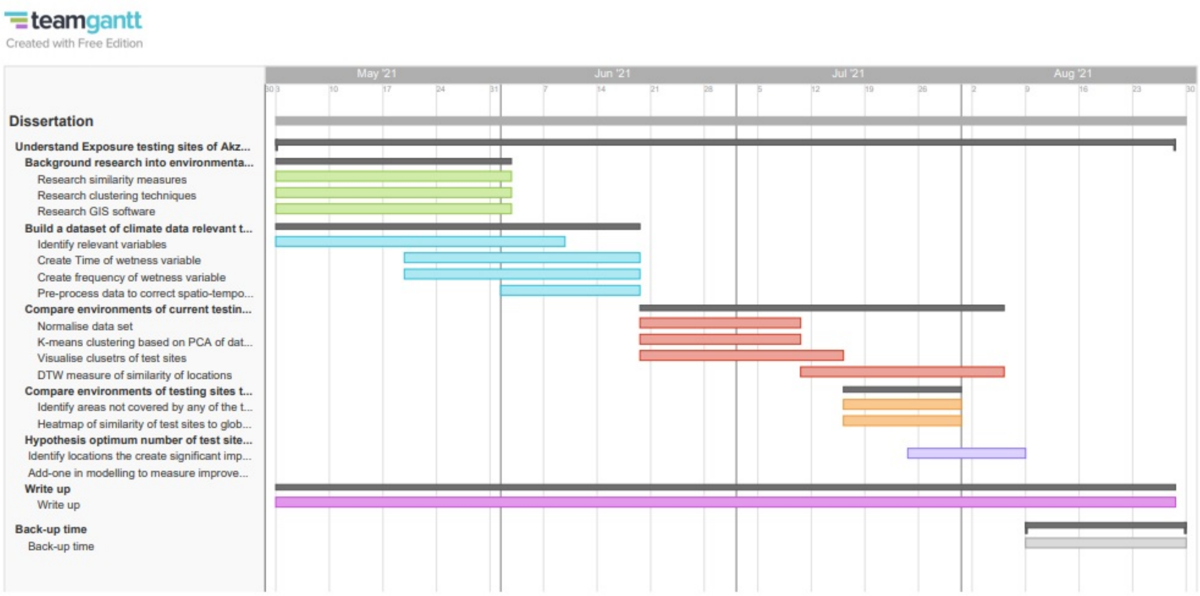


Figure 2: Project plan Gantt Chart

## References


C. Daly. Guidelines for assessing the suitability of spatial climate data sets. *International Journal of Climatology*, 26(6):707–721, 2006. ISSN 08998418. doi: 10.1002/joc.1322.

D. de la Fuente, I. Díaz, J. Simancas, B. Chico, and M. Morcillo. Long-term atmospheric corrosion of mild steel. *Corrosion Science*, 53(2):604–617, 2011. ISSN 0010938X. doi: 10.1016/j.corsci.2010.10.007. URL `http://dx.doi.org/10.1016/j.corsci.2010.10.007`.

ECMWF. CMIP5 daily data on single levels, 2016. URL `https://cds.climate.copernicus.eu/cdsapp#!/dataset/projections-cmip5-daily-single-levels?tab=overview`.

globenewswire.com. Industrial coatings market to reach 103.2 bn by 2025, 2020. URL `http://www.globenewswire.com/en/news-release/2020/10/08/2105977/0/en/Industrial-Coatings-Market-Size-to-Hit-US-107-32-Bn-by-2027.html`.

B. Gönençgil. Evaluate Turkey's Climate Classification by Clustering Analysis Method. pages 41–53. Springer International Publishing, Cham, 2020. ISBN 978-3-030-28191-5. doi: 10.1007/978-3-030-28191-5_4. URL `https://doi.org/10.1007/978-3-030-28191-5_4`.

K. Kozjek, M. Dolinar, and G. Skok. Objective climate classification of Slovenia. *International Journal of Climatology*, 37(March):848–860, 2017. ISSN 10970088. doi: 10.1002/joc.5042.

N. LeBozec, D. Thierry, P. Le Calvé, C. Favennec, J. P. Pautasso, and C. Hubert. Performance of marine and offshore paint systems: Correlation of accelerated corrosion tests and field exposure on operating ships. *Materials and Corrosion*, 66(3):215–225, 2015. ISSN 15214176. doi: 10.1002/maco.201307340.

N. LeBozec, D. Thierry, and K. Pelissier. A new accelerated corrosion test for marine paint systems used for ship's topsides and superstructures. *Materials and Corrosion*, 69(4):447–459, 2018. ISSN 15214176. doi: 10.1002/maco.201709814.

P. Netzel and T. Stepinski. On using a clustering approach for global climate classification. *Journal of Climate*, 29(9):3387–3401, 2016. ISSN 08948755. doi: 10.1175/JCLI-D-15-0640.1.

S. Park, H. Park, J. Im, C. Yoo, J. Rhee, B. Lee, and C. G. Kwon. Delineation of high resolution climate regions over the Korean Peninsula using machine learning approaches. *PLoS ONE*, 14(10): 1–23, 2019. ISSN 19326203. doi: 10.1371/journal.pone.0223362.

J. P. Praene, B. Malet-Damour, M. H. Radanielina, L. Fontaine, and G. Rivière. GIS-based approach to identify climatic zoning: A hierarchical clustering on principal component analysis. *Building and Environment*, 164(July):106330, 2019. ISSN 03601323. doi: 10.1016/j.buildenv.2019.106330. URL `https://doi.org/10.1016/j.buildenv.2019.106330`.

A. Rosenberg and J. Hirschberg. V-Measure: A conditional entropy-based external cluster evaluation measure. *EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (June):410–420, 2007.

E. Schindelholz and R. G. Kelly. Wetting phenomena and time of wetness in atmospheric corrosion: A review. *Corrosion Reviews*, 30(5-6):135–170, 2012. ISSN 03346005. doi: 10.1515/corrrev-2012-0015.

J. Zscheischler, M. D. Mahecha, and S. Harmeling. Unsupervised clustering of geophysical data : A critical analysis of traditional climate classifications. *Procedia Computer Science*, 00:1–10, 2012.