**Big Climate Data Analytics: Effective Knowledge Discovery from Colombia’s Weather Data**

By

Roland Hudson

A DISSERTATION

Submitted to

The University of Liverpool

in partial fulfilment of the requirements

for the degree of

MASTER OF SCIENCE

24/11/2018

ABSTRACT

**Big Climate Data Analytics: Effective Knowledge Discovery from Colombia’s Weather Data**

By

Roland Hudson

The goal for the project is a big climate data analytic system that enables knowledge-discovery to provide recommendations for construction strategies given a geographical location and the associated weather data.

The major aim of the research is to develop a distributed machine learning application that classifies Colombia climate data and provides decision support to environmental designers seeking to understand the spatial and temporal use of low-energy design strategies. These strategies can help provide more comfortable living and working conditions for people using the buildings and reduce the need for heating and cooling, lowering emissions and energy consumption.

In Colombia the weather is massively varied due to high altitude mountains, coastlines and effects of phenomena such as el Niño. Tropical weather is unlike the weather in the Northern and Southern latitudes as there are no seasons instead daily variations dominate. Identifying appropriate design strategies for construction that respond to climate in Colombia is particularly challenging.

The research examines literature on low-energy design strategies, the spatio in Colombia

Recommended content:

* Introduce into the research or technology area addressed in the dissertation
* Identify specific problems and challenges identified in your work
* Describe what is achieved and proposed in the project
* Briefly describe the structure of the report
* Mention what is the application area and benefits of the proposed solutions

DECLARATION

[check with the recent templates]

I hereby certify that this dissertation constitutes my own product, that where the language of others is set forth, quotation marks so indicate, and that appropriate credit is given where I have used the language, ideas, expressions, or writings of others.

I declare that the dissertation describes original work that has not previously been presented for the award of any other degree of any institution.

Signed,

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

Contents

[Chapter 1 Introduction 5](#_Toc528253156)

[1.1 Scope 5](#_Toc528253157)

[1.2 Problem statement 5](#_Toc528253158)

[2.1.1 Complexity of existing workflow 6](#_Toc528253159)

[1.3 Approach 6](#_Toc528253160)

[3.1.1 Literature review 6](#_Toc528253161)

[3.2.1 Artefact 7](#_Toc528253162)

[3.3.1 Application development methodology 7](#_Toc528253163)

[3.4.1 Qualitative evaluation of artefact by domain experts: 8](#_Toc528253164)

[3.5.1 Quantitative evaluation of output from artefact results: 8](#_Toc528253165)

[1.4 Outcome 8](#_Toc528253166)

[Chapter 2 Background and review of literature 9](#_Toc528253167)

[2.1 Background 9](#_Toc528253168)

[2.2 Literature review 9](#_Toc528253169)

[2.1.2 Low Energy Environmental Design Strategies 9](#_Toc528253170)

[2.2.2 Climate data is spatiotemporal 11](#_Toc528253171)

[2.3.2 Examples of applied ST data mining methods applied to climate data 11](#_Toc528253172)

[2.4.2 Clustering applied to climate classification 12](#_Toc528253173)

[2.5.2 Challenges for ST data mining 15](#_Toc528253174)

[2.6.2 Cloud computing 15](#_Toc528253175)

[2.7.2 Big data tools 16](#_Toc528253176)

[2.3 Summary 19](#_Toc528253177)

[3.1.2 Data: 19](#_Toc528253178)

[3.2.2 Analytics: 19](#_Toc528253179)

[3.3.2 Expertise: 20](#_Toc528253180)

[Chapter 3 Analysis and Design 20](#_Toc528253181)

[3.1 System actors 20](#_Toc528253182)

[3.2 General Use Case Analysis 21](#_Toc528253183)

[2.1.3 Define workflow 21](#_Toc528253184)

[2.2.3 Run workflow + monitor resources 22](#_Toc528253185)

[2.3.3 Output + visualise results 22](#_Toc528253186)

[2.4.3 Manage design strategies 23](#_Toc528253187)

[2.5.3 Proposed System Architecture 23](#_Toc528253188)

[3.3 Sequence Diagrams for General use cases 25](#_Toc528253189)

[3.4 Design methodology 27](#_Toc528253190)

[Chapter 4 Implementation 28](#_Toc528253191)

[4.1 Development iterations 28](#_Toc528253192)

[4.2 Environment 28](#_Toc528253193)

[4.3 Architecture 29](#_Toc528253194)

[4.4 Packages 30](#_Toc528253195)

[4.5 Climate data 32](#_Toc528253196)

[4.6 Workflow management system 33](#_Toc528253197)

[6.1.4 Workflowbuilder and Coordination package class diagrams 33](#_Toc528253198)

[6.2.4 Workflow creation 34](#_Toc528253199)

[6.3.4 Monitoring resources and workflows 36](#_Toc528253200)

[6.4.4 Running a workflow 37](#_Toc528253201)

[6.5.4 Management of design strategies 39](#_Toc528253202)

[4.7 Analytics 40](#_Toc528253203)

[7.1.4 Clustering Parameters and Filter Data 42](#_Toc528253204)

[7.2.4 Clustering Performance Indices 43](#_Toc528253205)

[7.3.4 Clustering classification and design strategy assignment 44](#_Toc528253206)

[7.4.4 Output of results 45](#_Toc528253207)

[4.8 Dashboard + visualisation 46](#_Toc528253208)

[8.1.4 Overview 49](#_Toc528253209)

[8.2.4 Cluster explorer 49](#_Toc528253210)

[8.3.4 Comfort comparison 51](#_Toc528253211)

[Chapter 5 Results and evaluation 52](#_Toc528253212)

[5.1 Evaluation 52](#_Toc528253213)

[1.1.5 Statistical comparison of the different knowledge discovery methods used 53](#_Toc528253214)

[1.2.5 Statistical comparison between individual knowledge discovery methods 53](#_Toc528253215)

[1.3.5 Opinion by domain experts 53](#_Toc528253216)

[Chapter 6 Conclusions 55](#_Toc528253217)

[6.1 Findings 55](#_Toc528253218)

[6.2 Knowledge discovery 55](#_Toc528253219)

[6.3 Evaluation results 55](#_Toc528253220)

[6.4 Future research 55](#_Toc528253221)

[Chapter 7 References 56](#_Toc528253222)

[Chapter 8 List of Tables 61](#_Toc528253223)

[Chapter 9 List of Figures 61](#_Toc528253224)

# Introduction

## Scope

This dissertation examines how a big weather data framework combined with knowledge discovery techniques can help define localized approaches to building design and construction that improve living conditions and reduce energy consumption in Colombia.

## Problem statement

The construction industry is responsible for massive amounts of greenhouse gases (Dimoudi and Tompa, 2008). Through their lifecycle buildings consume 50% of all total energy demand and contribute 50% of all CO2 emissions.A primary goal for environmental construction is to reduce the energy consumed by buildings, estimated to be around 40% of the total global energy consumption (Omer, 2008). Much of this energy is expended on heating lighting and cooling. Energy efficiency can be defined as the minimising the amount of energy consumed to achieve thermal comfort for occupants. Currently the energy required to maintain thermal comfort accounts for 60-70% of energy consumed in non-industrial buildings (Omer, 2008). Better understanding of human response to climatic context (bioclimatic design) can result in buildings that require less energy for heating and cooling (Olgyay and Olgyay, 2015, p11).

Low-energy construction strategies exist that can minimize or remove the need for heating and cooling in buildings. For example; orientation of buildings, sizing and positioning of openings, choice of materials and use of passive heating and ventilation. These techniques require an understanding of local and regional climate conditions across different time frames.

Detecting climate patterns in the Colombian context is challenging due to weather variations caused by extreme changes in altitude over short distances, proximity to different ocean bodies and phenomena such as el Niño and la Niña. This underlying complexity is complicated further by subtle and inconsistent seasonal patterns associated with tropical latitudes.

The Colombian context contrasts with Northern and Southern latitudes where consistent seasonal variation dominates the climate making it easier to identify what design strategy to apply. Colombia’s complex climate patterns make identifying appropriate, localized, low-energy construction strategies difficult. Typical construction is often unable to cope with regional and daily variations in weather, people live and work in uncomfortable conditions often too hot and too cold. To correct these issues heating and cooling systems are required which are expensive to install, costly to run and produce emissions.

### Complexity of existing workflow

For an architect or engineer the current workflow to determine what low-energy construction strategies to use is a multi-step approach:

* Analysis of climate data - analysis of historical weather data compiled into files representing typical meteorological years (TMY) with a range of variables stored for each hour of the year.
* Biological evaluation – data is plotted on a psychrometric chart (physical and thermal properties of moist air) to diagnose the hours of the year lying outside a predefined zone of thermal comfort.
* Identifying technological solutions (design strategies) – in response to the biological evaluation technological approaches can be identified to ensure more time within the comfort zone. For example; active or passive solar heating to gain heat energy from solar radiation.
* Developing the architectural application – the designer synthesises the previous three steps into a design proposal.
* Simulation may be undertaken to confirm the design approach or optimise a chosen strategy.

The designer must also consider usage patterns of the building increasing the complexity of the process. Buildings are rarely occupied constantly, depending on use, occupancy can vary daily (residential buildings are often occupied during evenings and night time), weekly (office buildings not in use at weekends) and through the year (schools and universities have seasonal holiday periods). The design approach for two buildings with different uses in the same location will not be the same. Usage patterns, activity types and orientation of spaces within buildings can also vary. Each space can therefore require a different design response.

To address these issues the project proposes an application based around Colombian weather data that combines datamining techniques with expertise of the low-energy building construction domain to link specific design strategies with a specific location and time frame.

## Approach

The project methodology begins with a literature review, then, following a specific development model, an IT artefact is developed. Finally, the artefact is evaluated qualitatively and quantitively.

### Literature review

The literature review surveys low-energy architectural design strategies and how these relate to weather and climate conditions and design criteria. The spatiotemporal nature of climate data is examined and techniques for data mining meteorological data are discussed. The review also includes big data tools, components, applications and architecture in climate science. With a focus on the use of workflow management for scientific big data systems and appropriate analytics methods. Knowledge discovery as a process model is defined and its application to pattern seeking with climate data is addressed.

### Artefact

The IT artefact is an application that facilitates big weather data analytics for Colombia. By integrating analytics and visualisation, the application enables data exploration and knowledge discovery by linking construction strategies with geographical location and related historical weather data.

### Application development methodology

Development follows a UML based agile model driven design method with a distribution of development phases and activities show in Figure 1. Each activity involves specific techniques and can be linked to deliverables and specific UML models (Table 1).

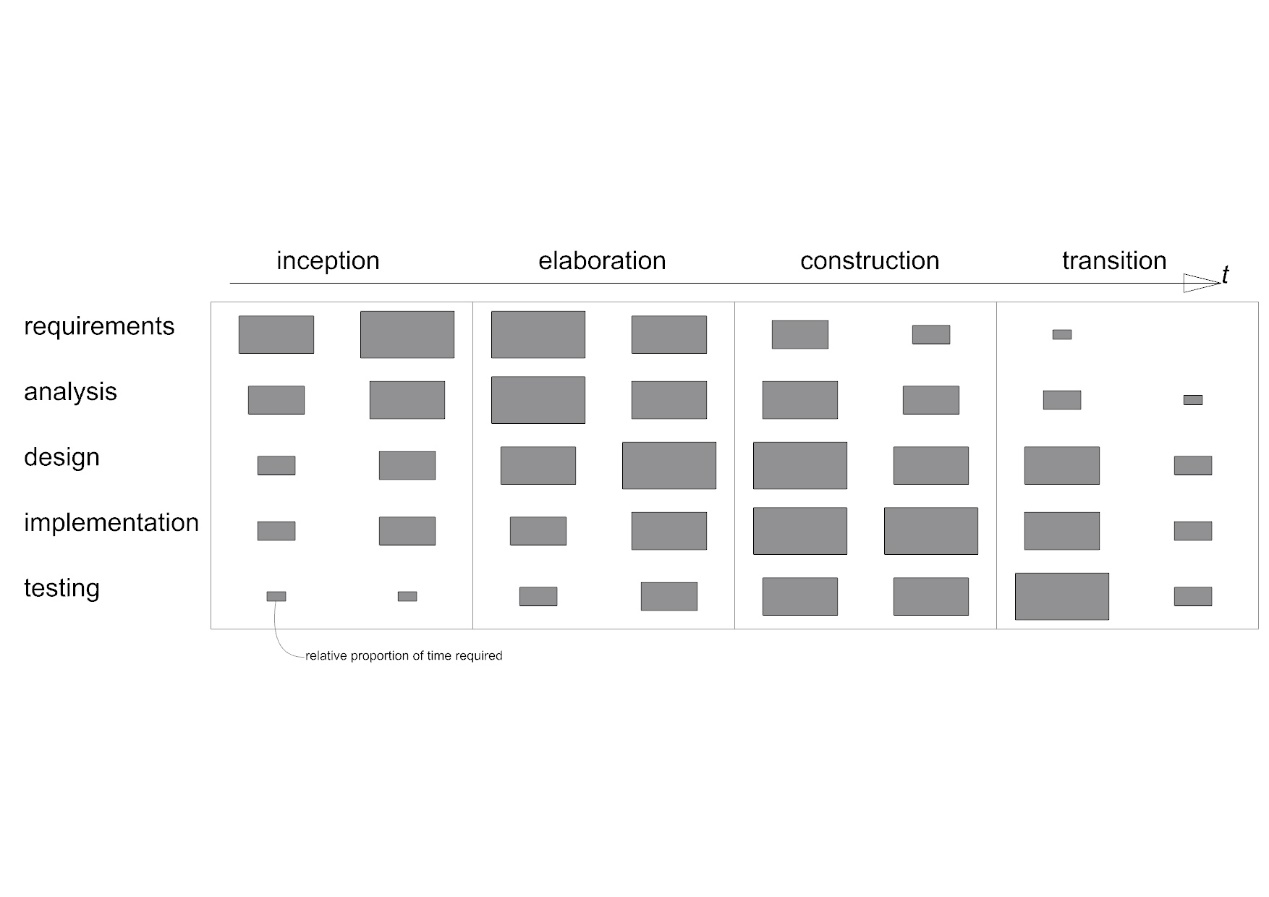
****

Figure 1 Agile model driven design phases

|  |  |  |  |
| --- | --- | --- | --- |
| **Activity** | **Techniques** | **Deliverables** | **Diagrams** |
| **Requirements capture and modelling** | Text descriptions of use cases and requirements  Use case modelling  Architectural modelling, prototypes | Use case model, requirements list, initial architecture | Use case, Package |
| **Requirements analysis** | Use cases analysed to extract required objects. Interactions between objects identified – communication diagrams developed | Analysis models | Class, Object, Communication |
| **System and architecture design** | Design patterns identified | Overview design and implementation architecture | Package, Component, Deployment, Class |
| **Class design** | Class and object modelling, Interaction modelling. State modelling, Design Patterns  prototypes | Design models | Class, Object, Sequence, State machine, Package |
| **Interface design** | Class and object modelling, Interaction modelling, State modelling, Design Patterns, prototypes | Design models, interface specification | Class, Object, Sequence, State machine, Package |
| **Data management design** | Class and object modelling, Interaction modelling, State modelling, Design Patterns, prototypes | Design models, data specification | Class, Object, Sequence, State machine, Package |
| **Construction** | Programming, component reuse | Constructed system, documentation |  |
| **Testing** | Programming, test planning and design, testing | Test plans, test cases, tested system |  |

Table 1 Development activity details

### Qualitative evaluation of artefact by domain experts:

Qualitative evaluation is based on software walkthroughs, presentations and interviews with domain experts. Experts will be presented with a series of studies and results from the application, opinions of experts will be captured and summarised. With expert review it will be possible to evaluate if useful localised construction approaches can be generated.

### Quantitative evaluation of output from artefact results:

Quantitative evaluation involves statistical comparison of different analytic methods, and different configurations of those methods. Quality measures for the methods applied include distance metrics for clustering. The application is also evaluated through verification, validation and testing – tests identified during the requirements specification and revisited through the prototyping stages.

## Outcome

The goal for the project is a big climate data analytic system that enables knowledge-discovery to support decision making in the design and construction of buildings in Colombia. Specifically, the project aims to enable the search for patterns in climate data that can be linked to localized, climate-responsive design and construction strategies. Application of these strategies can lead to buildings that perform better in terms of production costs, life-time running costs (reduced heating and cooling) and occupant comfort.

The project proposes that by combining a big data workflow management infrastructure with spatiotemporal data mining techniques localized approaches to building design and construction that respond to the unique weather conditions in Colombia can be identified.

To achieve this goal a big data system is developed that follows current best practices for the storage, processing, analysis, management and visualization of the data. Specific focus will be on enabling the analytics and visualization to facilitate knowledge discovery through data mining. Knowledge from the data will support decision making for the design and construction of buildings to potentially improve living conditions (quality of life and wellbeing) and reduce energy consumption in buildings.

# Background and review of literature

## Background

The literature review spans various themes, first low-energy environmental design strategies are examined, what they are and how they are represented is defined. The spatiotemporal nature of climate data is described, this is contrasted with classical data mining and key differences are determined. Data mining methods applicable to spatiotemporal are discussed and challenges for these techniques identified. Data mining is positioned as one of three key elements in knowledge discovery processes, the relationship with the other two; domain expertise and data management is described.

Relevant big data concepts are explored, this includes the use of workflow management systems for big data science applications and how Infrastructure as a Service offered by cloud service providers may be an applicable service model. Applications that have addressed climate data tasks using big data analytics tools are identified. Common to many of these applications is Apache Spark’s machine learning library, this is identified as a key tool for this project.

## Literature review

### Low Energy Environmental Design Strategies

Human thermal comfort Can be understood as a combination of temperature, relative humidity, air movement and radiant temperature, giving a state of mind where a person requires no change in current conditions (ASHRAE, 2013) or a state where minimal extra energy is require to maintain the human balance (Manzano-Agugliaro *et al.*, 2015).

Psychrometric charts are used to map interrelationships of thermal conditions of the environment (Figure 2). A zone of human thermal comfort can be plotted following standard guidelines (ASHRAE, 2013). Hourly data points can be plotted on the chart, where they fall outside of the comfort zone the design of the building and/or services must be adapted to provide comfort.

The psychrometric chart was adapted by Givoni (1992) as the building bioclimatic chart and later Manzano-Agugliaro *et al.* (2015) to include specific zones representing strategies that can be applied to a design to extend the zone of comfort. Conventional heating and air conditioning are recommended only at extremes. The strategies include what Lechner (2009, p9) describes as tier 1: basic building design (building orientation, position and size of openings, material specification) and tier 2: passive systems (passive solar heating, night-time flush cooling). Correct design decision making at these levels can reduce building energy consumption by up to 80% (Lechner, 2009, p9).

Each design strategy can be further specialised into a series of instrumental techniques (Manzano-Agugliaro *et al.,* 2015) which can include regionally specific and traditional construction and more experimental methods. Climate Consultant software (Milne, Liggett and Benson, 2009) generates a prioritised subset of recommendations from a set of 68 design guidelines each associated with a zone on the psychrometric chart.

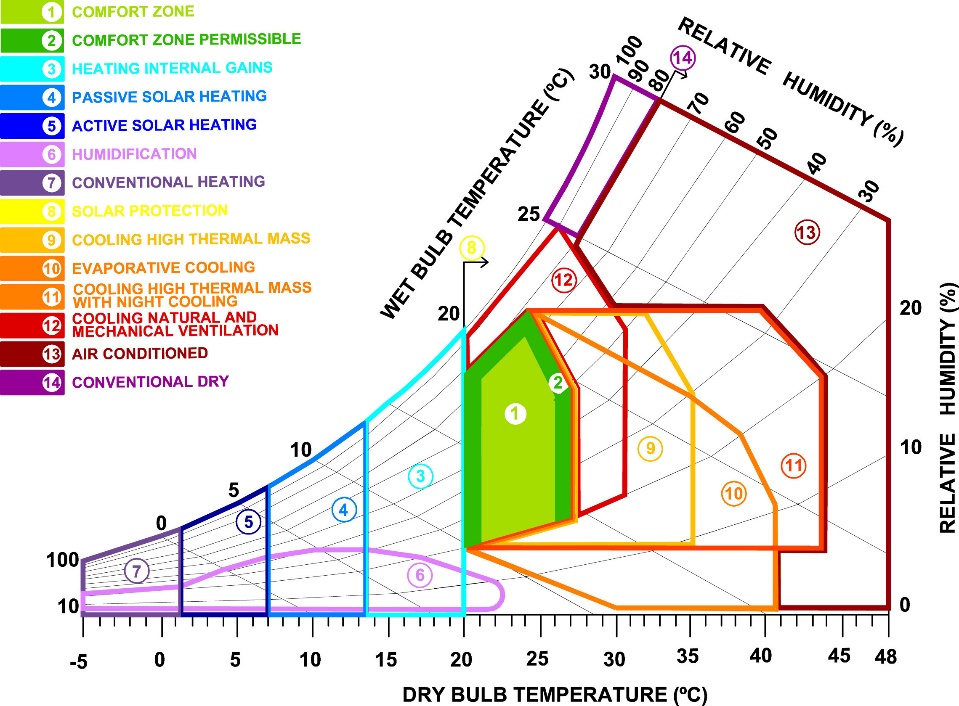


Figure 2 Manzano-Agugliaro *et al.* (2015) adapted version of the psychrometric chart

#### Comfort Indices

Colombia’s national institute of hydrological, metrological and environmental studies (INSTITUTO DE HIDROLOGIA, METEOROLOGIA Y ESTUDIOS AMBIENTALES (IDEAM)) has proposed a method to calculate climatic comfort (IDEAMCI) in Colombia (González, 1998). Based on a system (by unreferenced authors Leonardo Hill and Morikofer – Davos) for refrigeration power it assumes when the human body is surrounded by air temperatures lower than 36.5C and the air is constantly renewed by wind a cooling effect will be experienced. IDEAMCI extends this by including relative humidity at three discrete ranges of altitude and produces an index on a scale of zero to fifteen or more (hot to cold). IDEAMCI seeks to resolve an understanding at both a spatial and temporal resolution (hourly) across an entire Country.

Most recently the Universal Thermal Comfort Index (UTCI) has been proposed citing the shortcomings of preceding indices and limits in terms of confinement to specific applications (Jendritzky and Höppe, 2017).

Indices exist to help those not familiar with the domains of thermos-physiology or biometeorology better understand the implications of climate, activity and clothing on the human body (Höppe, 1999) and are not necessarily designed to support decision making for construction. Notably UTCI is the result of a multi-disciplinary effort the development team did not include practitioners from the built environment. Indices are useful in later design stages to fine tune heating, cooling and ventilation systems or to refine material specifications (insulation and glazing) ensuring certain thermal comfort levels are achieved. Indices convert a multivariate problem into a single value removing the visibility of the underlying data making it difficult to determine design strategies that respond to the cause of discomfort.

Architects concerned with designing comfortable environments need to be aware of the full set of climate variables early in the design process to develop appropriate strategies. Strategies inform siting and orientation of the building, size and number of openings, materials and integration of technological systems to assist passive cooling and heating. Such early stage design decisions are more probable to have an impact on the cost and function of the finished building and are cheaper to implement (CURT, 2004, p4). Despite their short comings these indices can be helpful to compare and evaluate clustering solutions.

### Climate data is spatiotemporal

Data collected for climate science is classed as spatiotemporal (ST) data (Atluri, Karpatne and Kumar, 2017). Mining and knowledge discovery with ST data differs from classical data mining due to its properties and the variety of data types (Faghmous and Kumar, 2014) . ST data is heterogenous, it is not identically-distributed, instead ST data demonstrates non-stationarity in space and time. Auto-correlation exists in ST data two nearby locations are not independent but are correlated. ST data can be categorised into four types; event data (start and end of heavy rainfall), trajectory data (path of a cyclone), point data (temperature measured in a moving set of weather balloons) and raster data (temperature measured across a fixed set of weather stations).

Classical data mining uses features with labels (Atluri, Karpatne and Kumar, 2017) in ST instances can be defined as points, trajectories, time-series, spatial maps and raster. Search for similarities between these instances involves clustering, classification, pattern discovery and relationship mining.

Various data mining methods applicable to climate data are described by Atluri, Karpatne and Kumar (2017). *Relationship mining* involves linking changes in one variable to other phenomena. *Frequent pattern mining* includes searching *motif patterns* in time-series and in networks for sets of distant locations experiencing similar climatic conditions with consistent temporal activity. Change detection can identify transitions or deviations in behaviour. Faghmous and Kumar (2014) add *network-based analysis* to undertake relationship and pattern mining in gridded and non-gridded climate data sets. *Pattern mining* for Faghmous and Kumar includes searching for user defined patterns, empirical orthogonal functions and ST clustering.

### Examples of applied ST data mining methods applied to climate data

* Self-organising maps (SOM)’s have been used to extract features from data (Liu, Weisberg and Mooers, 2006) and applied to metrology and oceanography (Liu and Weisberg, 2011)(Liu and Weisberg, 2005).
* Delta-maps (Fountalis, Bracco and Dovrolis, 2014) (Bracco *et al.*, 2017) group nodes in a network according to homogeneity, these have beenapplied to precipitation and sea surface temperatures. Robustness analysis of networks generated can be evaluated using link maps, area strength and s-core decomposition.

### Clustering applied to climate classification

Recent research suggests that clustering, an unsupervised learning technique, is particularly applicable to climate analytics, classification and data mining. Studies in the literature focus on common clustering themes; seeking best performing clustering methods for a specific goal, comparisons between hierarchical and non-hierarchical clustering methods, selecting the number of clusters, pre-processing variable prior to clustering and developing hybrid workflows from different clustering methods. The methods described in the following sections work with vectors, each a set of *n* features. Within this n-dimensional parameter-space Euclidian distances can be calculated between vectors and between vectors and cluster centroids to determine proximity, which is used to define which vectors are in which clusters.

#### Hierarchical clustering

Hierarchical clustering (HC) methods organise data into a sequence of nested groups. HC begins by treating all observations of vectors as individual clusters. Two steps are repeated first identify the two closest clusters and then merge these into a single cluster Figure 3. Fovell & Fovell (1993) studied HC methods to define climate zones in the US and found average linkage method performed best to minimise bias in terms of method, latent and information.

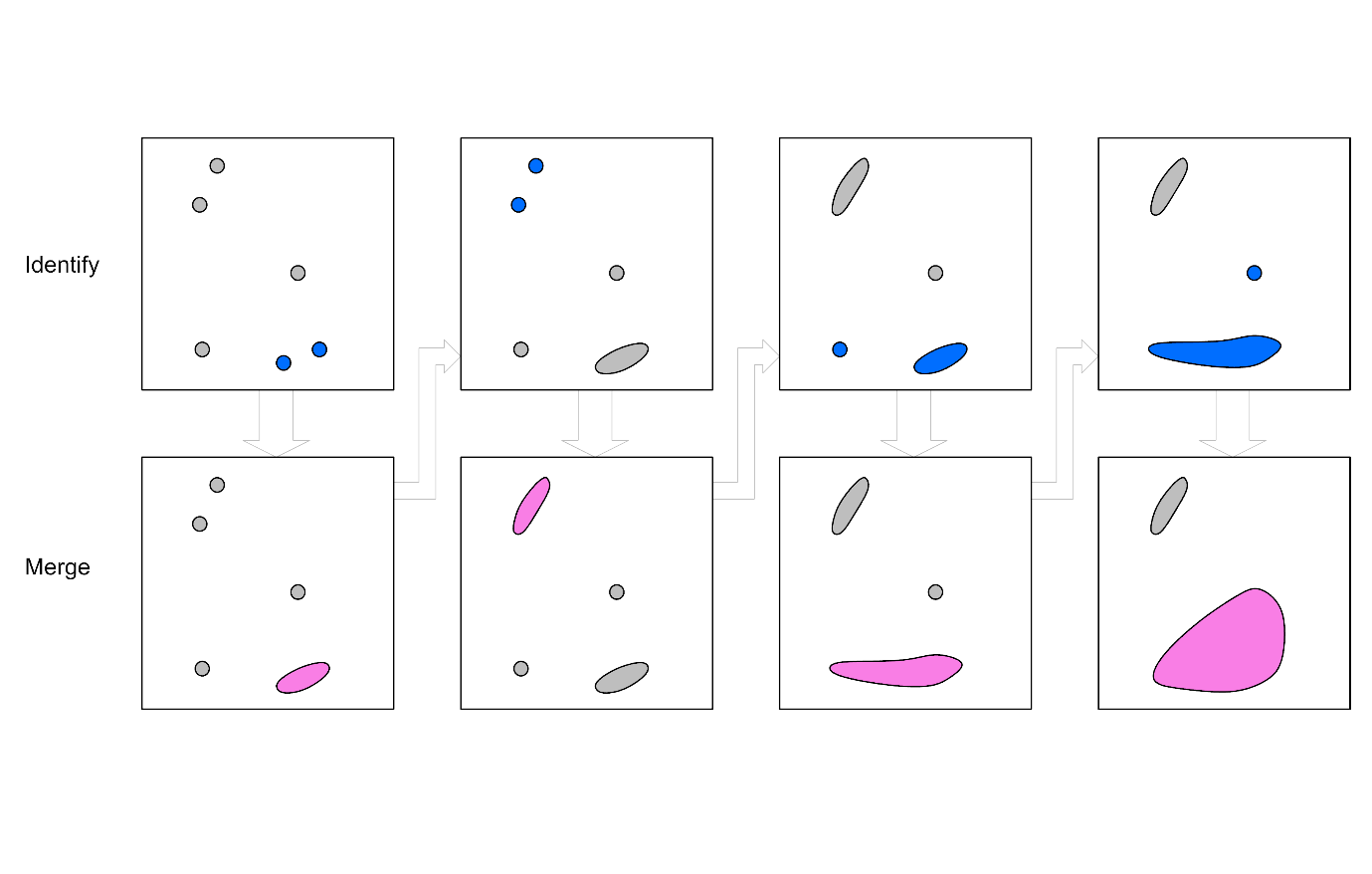


Figure 3 Hierarchical clustering steps

#### Non-hierarchical clustering

Non-hierarchical clustering (NHC) methods provide a single partitioning of the data aiming to define natural groups within the data (Jain and Dubes, 1988, p89). K-means is a NHC method that was defined in 1955 (Jain, 2010), despite its age and many alternatives it remains popular in the literature of clustering with climate data. The k-means algorithm requires *k,* the number of clusters to be specified and begins by creating *k* initial cluster centroids. Two steps are repeated; vectors are assigned to a cluster with the closest centroid, cluster centroids are redefined based on the data points in the cluster.

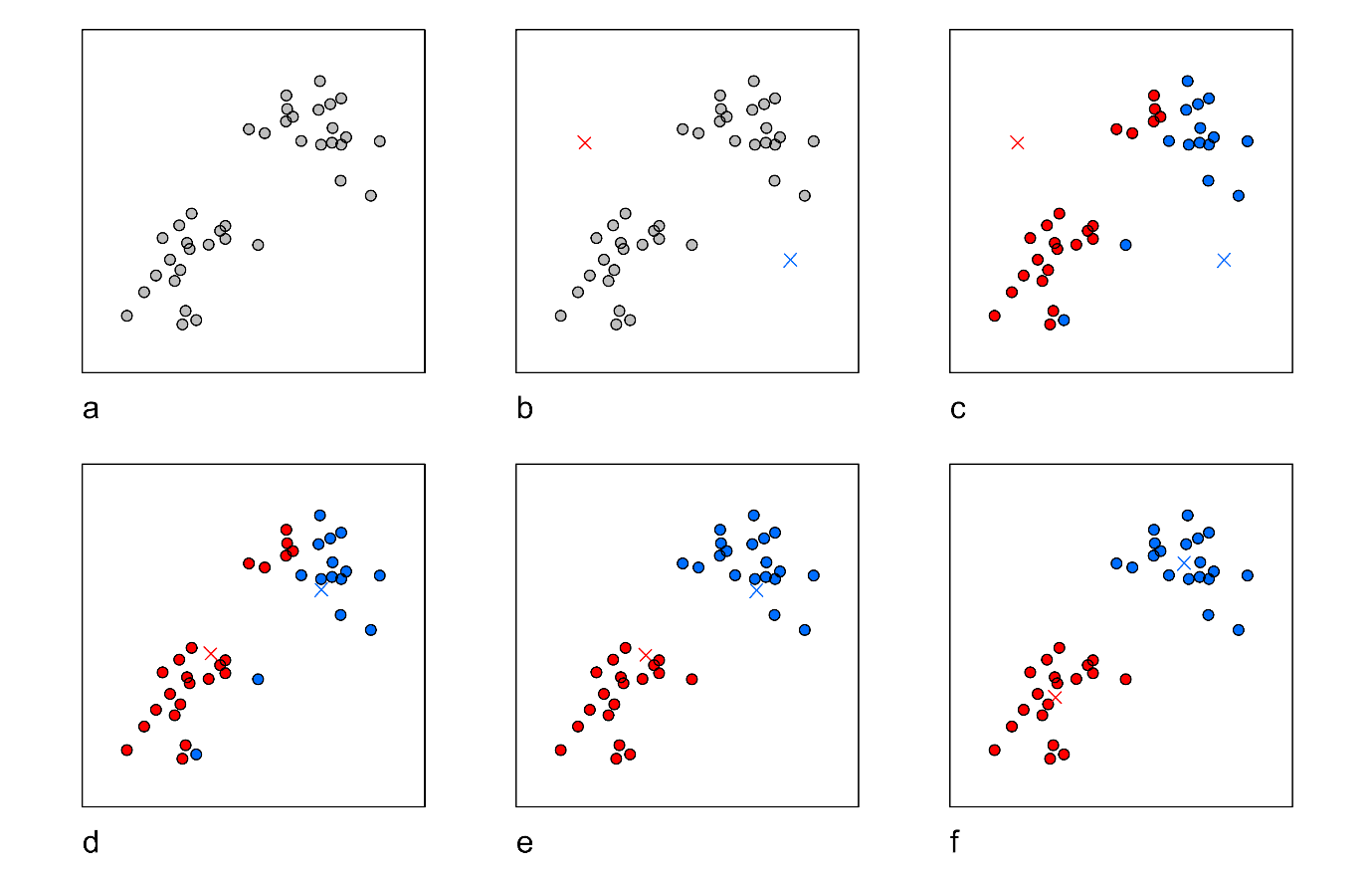


Figure 4 Non-hierarchical k-means steps: (a) Original data. (b) Random initial centroids. (c-f) Two iterations of: assign data points to closest cluster, recalculate cluster centroid. Adapted from Piech (2013).

K-means NHC was compared to Ward’s HC technique (Degaetano, 1996) in a study that sought to develop an ecosystem management and planning guide by defining mesoscale climate zones in the north-eastern US. K-means was found to improve the spatial distribution, homogeneity of clusters and produced stable clusters with minimal information bias.

Clustering was used to assess climatic influences on water resources and food security in the Himalayas (Forsythe, Blenkinsop and Fowler, 2015). It showed clustering could be used to characterise the bias of gridded datasets and undertake meteorological reanalyses of climate models. K-means defined eight sub-regional climate classifications and further increases in cluster numbers defined subdivisions to each macro zone.

Zscheischler, Mahecha and Harmeling (2012) used k-means clustering with subsets of five normalized variables. When k-means was used with climate and vegetation variables similar clusters to the Köppen Geiger Climate Classification (KGCC) zones could be generated.

Netzel and Stepinski (2016) undertook 32 different clustering experiments comprised of HC Ward linkage or NHC k-means with different input variables and similarity measures. These were compared to KGCC. The study concluded that clustering could find 50% of the climate types defined by the KGCC. The remaining classes differed in climatic character and spatial distribution but were shown to be more homogeneous and more distinct than KGCC types. The study concluded than indicated that the NHC k-means gave better results than the HC.

#### Hybrid approach

Rhee *et al.* (2008) used k-means as part of a multi-step approach to delineate climate regions in the Carolinas that combined in-situ (weather station data) with remotely sensors and spatially distributed data. K-means was integrated within a more complex workflow (hierarchical clustering followed by non-hierarchical then decision trees were trained on results that classified remotely sensed data) and the study demonstrated the validity of the method for establishing clusters that were subsequently used for supervised classification of data.

#### Choosing the numbers of clusters

A recurring theme in the literature is how to choose the number of clusters, *k*. Many indices exist to describe the homogeneity and completeness of clustering solutions (Arbelaitz *et al.*, 2013), but Fovell and Fovell (1993) insist informed decisions over *k* must also include domain knowledge. Increasing *k* serves to become more specific but at the cost of generality. Lower cluster numbers represent a loss of detail but, they can enhance interpretation and generality. Climate data varies smoothly and hard edges between clusters do not exist, the choice is partly subjective and based on an adequate subdivision.

Silhouette index represents dissimilarity (Wiwie, Baumbach and Röttger, 2015) and is found to produce good results in comparison to other indices (Arbelaitz *et al.*, 2013). Silhouette is the ratio between the difference of the inter-mean (average of distances from each cluster centroid to the global centre) and the intra-mean (the average distance between each vector and its cluster’s centroid) and the maximum

Dunn index defines quality of clustering as a ration of cohesion (as the maximum cluster diameter) and separation (the distance to the nearest neighbour. Luna-Romera *et al.*, (2016) suggest the first maximum found with the Dunn index indicates the optimal number of clusters. Higher values in both Silhouette and Dunn indices indicate better performance.

Sum of Squared Errors (SSE), a measure of internal clustering cohesion (Thinsungnoen *et al.*, 2015) can also be used for validity and is calculated using the square of the Euclidian distance between each vector and its cluster centroid. TheElbow method can be used to find a close to optimal value of *k* by plotting SSE against the *k.* The inflexion (elbow) in the graph indicates a good candidate for *k* (Kodinariya and Makwana, 2013 and Nikolaou *et al.*, 2012)

#### Pre-processing and choosing the variables

Netzel and Stepinski (2016) and Zscheischler, Mahecha and Harmeling (2012) are clear that normalisation of variables is essential to remove the effects from different scales. Principle component analysis (PCA) is thought to be important in identifying and removing highly corelated variables (Fovell and Fovell, 1993), but simultaneously questioned. PCA required subjective decisions as to where the truncation of the features should occur and lead to information bias. Rhee *et al.* (2008) chose to avoid of truncation entirely and not use PCA fearing loss of information in their monthly time series data.

Domain knowledge also needs to factor in selection of variables. Forsythe, Blenkinsop and Fowler (2015) suggest the usefulness of classification is dependent on the extents that it reflects the constraints that determine the physical processes of interest. This indicates that understanding the domain and existing common approaches to the problem can begin to justify which variables are required.

### Challenges for ST data mining

Dealing with interdependencies at multiple scales within climate data is complex and means global studies cannot be used to understand long-term local impacts (Faghmous and Kumar, 2014). Relationships in climate data may be long range and multivariate, many space-time-variable subsets exist where relationships may be found. This spatiotemporal variability makes clustering with ST data challenging. Similarly, anomalies and extremes in climate data need to be understood as multivariate cumulative extremes. Faghmous and Kumar suggest better methods are needed for validation of ST data mining. Significance testing needs randomization tests that do not break the inherent autocorrelation and performative measures are required to compare unsupervised STDM.

For Atluri, Karpatne and Kumar (2017) key challenges involve finding methods for combining multi-modal data sets and controlling granularity of partitioning to ensure substructures are not overlooked. The need to integrate domain theory and expertise is acknowledged (Karpatne *et al.*, 2017) as a key strategy that could accelerate knowledge discovery in data science particularly where complex physical phenomena are involved.

### Cloud computing

Cloud computing is defined by NIST (Mell and Grance, 2011) as a model consisting of the following characteristics; on demand service, board network access, rapid elasticity and a measured service. Three key service models are offered by cloud providers; Software as a Service, Platform as a Service and Infrastructure as a service (IaaS). IaaS provides access to cloud-based computing resources that allow the deployment and execution of arbitrary software. This service model offers three key functions of interest to this dissertation; dynamic provisioning and configuration of processing resources to run cloud-based systems, scalable storage capacity that can be used for applications, backups, archival, and file storage and Content Delivery Networks to store content and files to improve the performance and cost of delivering content for web-based systems (Liu *et al.*, 2011).

Amazon Web Services (AWS) is a cloud provider offering all the different cloud service models. The AWS IaaS includes Elastic Map Reduce (EMR) a hosted Hadoop framework that includes Apache Spark and other distributed frameworks. EMR automatically configures another Amazon product, Elastic Compute Cloud (EC2) to provide virtual cloud-based servers. With EMR clusters of virtual machines are launched on a Virtual Private Cloud (VPC). Specific analytic jobs are described as a Step on AWS, which is a distinct work unit that can run on an EMR cluster and a single cluster can have several Steps. EMR is designed to handle node provisioning, Hadoop configuration, cluster setup and tuning and automatically replaces poorly performing machine instances. EMR provides a file system, EMRFS that allows reading and writing files to AWS’s Simple Storage Service (S3). S3 is an object storage service providing scalability and automatic data replication by distributed data across a minimum of three facilities in a region.

### Big data tools

#### Workflow management for scientific big data systems

An approach for supporting scientific data analysis on large data sets in the cloud is workflow management systems (WMSs) (Buyya *et al.*, 2016). These processing tools enable acquisition of resources, scheduling of tasks, execution of data analysis and visualisation on distributed resources. Workflows are defined as a series of linked tasks in the form of directed acyclic graph (DAG).

Specific WMS platforms exist (see Askalon, Kepler, Taverna and Pegasus) some provide a graphical interface to assemble workflows costing of loops, conditionals and graph constructs. Originally developed for grid computing many have been extended to take advantage of the cost-effectiveness of cloud platforms and applicable to climate science (Figure 5) (Rodriguez and Buyya, 2017).

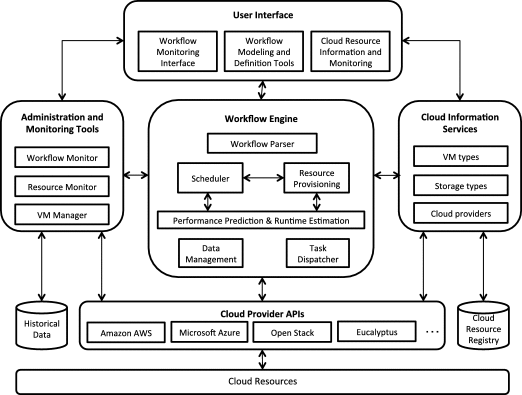


Figure 5 Reference architecture of a WMS

#### Big data application architecture and components

Avci Salma, Tekinerdogan and Athanasiadis (2017) define a big data feature model and a generic reference architecture that can be developed into an application architecture using domain-driven design. Features and architecture components can be selected based on design rules determined by the domain (Figure 6).

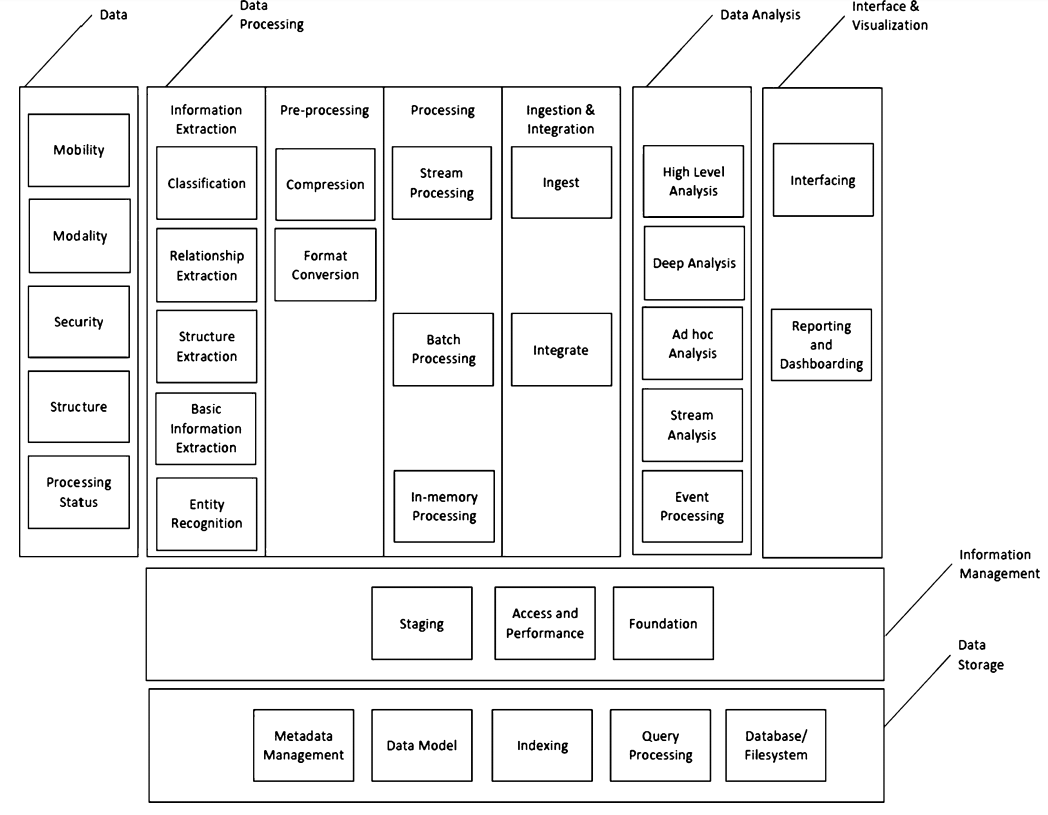


Figure 6 Big data reference architecture (Avci Salma, Tekinerdogan and Athanasiadis, 2017)

#### Big data and Weather Data

Several precedents exist describing the application of big data tools to process and undertake simple analysis on climate and weather data. Three tools frequently occur in the literature Hadoop, Apache Spark and MapReduce.

* Apache Spark was used to process at-rest weather data to determine averages for a range of variables across a set of weather stations (Jayanthi and Sumathi, 2017).
* Hadoop and Spark were used to find points with similar weather conditions using Euclidean distances based on weather attributes using the US National Oceanic and Atmospheric Administration (NOAA) hourly land based data set (Rodenburg and Maria Fiore, 2017).
* MapReduce was compared to Spark to calculate minimum, maximum and average values of weather parameters using data from NOAA (Chouksey and Chauhan, 2017). For smaller datasets differences in performance were less pronounced, above 8GB Spark was faster.
* Hadoop and MapReduce were used on NOAA data (Dagade *et al.*, 2015) to compare the performance of Pig and Hive to average data for each station for a single variable.
* MapReduce was used to build an analytical engine for processing National Climatic Data Center (NCDC) temperature data from automated sensors (Mariam Varghese, 2015).
* K-means with MapReduce was used to cluster weather data from China (Fang *et al.*, 2014). Different size datasets were tested to evaluate processing speeds (250mb-2GB) and compared against another clustering algorithm.
* A self-organising map (SOM) (a type of artificial neural network trained using unsupervised learning) was implemented using Apache Spark and analysed IoT data and found to reduce processing time compared to a serial method (Jayaratne *et al.*, 2017).
* MapReduce enabled prediction using an artificial neural network and k-means clustering with air quality data stored with HBase, in an implementation designed to support decision making in traffic regulations in Marrakesh (El Fazziki *et al.*, 2015).

#### Apache Spark

Apache Spark is a distributed processing system and unified analytics engine designed for processing large datasets.

Key features of Spark (Zaharia *et al.*, 2013) are:

* fault tolerance through a parallel recovery mechanism
* tolerance of stragglers (slow nodes)
* a processing model based on discretised streams (D-Streams)
* integration with batch and interactive query models such as MapReduce
* dynamic load balancing (Das, Zaharia and Wendell, 2015)

Fault recovery in other streaming systems is based on times data replication which is expensive in terms of time and hardware and results in long recovery times and problems for handling stragglers. Systems such as Storm, TimeStream and MapReduce use continuous operators that receive each record, internal states are updated, and new record sent. The Spark Streaming / D-Streams model contrasts with the long-lived operators approach by structuring the computation as a sequence of stateless, batch processes issued at short time intervals. Resilient Distributed Datasets (RDDs) are used to keep data in memory and by tracking the graph of operations used to produce each RDD it can be recovered without on disk replication. Faults on nodes are handled using parallel recovery which means on failure of a node all other nodes in the cluster work to rebuild the lost RDDs.

Spark’s Machine Learning Library MLlib offers several different clustering methods including k-means and bisecting k-means. Both methods provide access to a clustering “cost” the *within set sum of squared errors* (WSSSE) (sum of squared distances of points to their nearest centre) (*KMeansModel (Spark 2.3.2 JavaDoc)*, no date). Luna-Romera *et al.* (2016) suggest that Spark’s WSSSE does not capture cluster consistency or distance between clusters. As alternatives Spark based implementations of the Dunn index and Silhouette coefficient are described.

#### Challenges for big data analytics applied to climate science and construction

The use of big data analytics is underutilised in the domain of weather data (Jain and Jain, 2017), better big data weather forecasting could benefit the construction industry by helping build more efficient buildings and provide an “*improved quality of life”.*

#### Knowledge discovery

Knowledge discovery (KD) integrates data mining, domain theory and data management. KD is defined as (Begoli and Horey, 2012) 1. Collection, storage and organisation of data. 2. Understanding and application of analytic methods. 3. Understanding the problem domain. To enable knowledge discovery Begoli and Horey recommend applications are made to allow researchers easy ways to interact, explore and analyse data. A variety of analysis methods should be supported inclung statisitical, data mining, machine learning, visualisation and visual analysis. Different data storage and processing mechanisms should be provided to support a variety of intermediate data structures (structured and semi-structured) required by different ananlysis methods. Data should be made as accessible as possible by using open standards, lightweight architecture and APIs to expose results. The potential for knowledge discovery in climate science has not yet been fully realised (Bracco *et al.*, 2017).

## Summary

The three integrated elements that define the concept of Knowledge Discovery these provide a concise framework for defining goals of the dissertation.

### Data:

Key considerations involving the data include the growing quantities of climate data and its spatiotemporal nature. Large amounts of this form of data are available and the sources are growing which suggests the logic in seeking a solution that is scalable in terms of data capacity. AWS’s IaaS provides flexible, storage for data in S3, this service is included in AWS’s free usage tier and has a well-documented Java SDK. Understanding the spatiotemporal nature of climate data is important for manipulating and organising climate data especially in the variety of ways patterns may exist.

### Analytics:

The literature suggests that clustering is a key machine learning technique for datamining spatiotemporal data. Spark emerges as a main contender for undertaking this kind of climate data analysis, it offers NHC k-means and HC bisecting k-means as built in functions. Spark provides methods for evaluating the clustering performance and further validation metrics with Spark implementations were identified. Spark is included with EMR, AWS’s analytics IaaS that provides hosting of virtual machines for distributed processing. EMR is part of AWS’s free usage tier and a well-documented Java SDK is provided.

### Expertise:

The domain includes human biology its relation to climate and the construction of buildings. Understanding domain goals and concerns will provide insight into the types of analytic processes that may be required. This suggest the concept of the workflow and a management system that is tailored to the domain but integrates with the chosen big data services is a fundamental goal. Understanding the relationship of low-energy design methods and construction techniques to the climate data is crucial for analytics methodology. Furthermore, how these can be represented and manipulated is important in enabling the connection between environmental conditions and how these may be adjusted with certain design strategies to try an achieve human comfort in buildings.

# Analysis and Design

## System actors

The primary actor in the system is an Environmental Designer (ED), an architect or engineer whose aim is to develop energy-efficient designs for buildings in specific geographical locations. The ED’s goal is to minimise the need for heating and cooling systems while maintaining thermal comfort for occupants reduce energy consumption.

To achieve these aims the ED needs to understand the interrelationship of occupant’s thermal comfort, topography, local weather conditions, annual solar path and larger scale climatic patterns and examining how these interact with the anticipated usage patterns of the building.

Equipped with this knowledge the ED can apply a hierarchy of design strategies to develop the design. The first level relates to designing the form and fabric of the building (for example the location, orientation, materials of the building). The next level involves identification of potential passive strategies (for example, designing direct solar heating and natural ventilation into a building). The final level is the specification of mechanical systems to heat and cool, potentially using renewable energy sources. The first two levels are the concern of this project.

Other actors identified are a system administrator and data scientist. The administrator’s role is the configuration of systems settings such as access and security. The data scientist working in the field of environmental design is also identified as a system user. In this case the use is like the ED but with additional goals of management of the datasets, developing and managing new analytical techniques, visualisation and statistical methods that the ED has access to.

## General Use Case Analysis

The ED’s goal is to run an analytic or data mining technique on a set of climate data to identify patterns that can be visualised and linked to with specific design strategies. An example analysis / data mining problem could be framed as:

*For the coastal regions of Northern Colombia what design strategies can be identified for afternoon weather conditions during the first three months of the year, using a k-means clustering approach. Visualise the clusters on a map and indicate the design strategies and considerations applicable to each cluster.*

To address this broad aim data storage, analytical processing and visualisation should be developed as a proof-of-concept using a public cloud-based infrastructure as a service (IaaS) to take advantage of the low-cost capital investment, flexible infrastructure, performance and the potential for collaboration.

The ED should be able to define this process via an online interface that allows the storage, editing and reuse of previously defined processes. The ED requires the ability to start, pause, stop and cancel the analysis process. Prediction of runtime and monitoring of workflow progress is also desirable. Results should be stored and accessible for online visualisation and statistical summary and comparison. Resulting graphics and summaries should be available for download in formats for use in excel and as high-quality images and vector graphics for use in reports once the process is complete.

This general use case can be subdivided into four phases:

### Define workflow

To define a process or work-flow (Figure 4) the ED must be able to specify a dataset or collection of datasets selected from a set of preloaded data. From the data the ED needs to define a geographical region or single point to study. The ED requires a high level of control of the temporal dimensions of the data. They will define the start and end dates and may need to specific that analysis takes place on recurring time-periods within the data (such as an afternoon in a specific season). The ED needs to specify which variables from the data to analyse. Once the dataset and spatial and temporal subset has been defined and the dimensionality specified, the ED will select an analytic or data mining technique from a set of predefined (and described) methods. At any point in this process the ED may need to save, save as and or edit the defined work-flow.

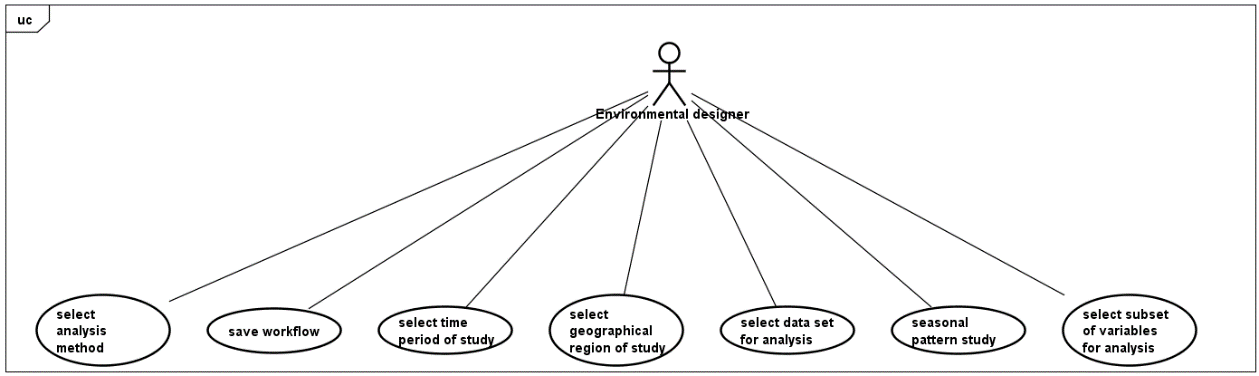


Figure 7 Define workflow

### Run workflow + monitor resources

Once the work-flow is defined the ED will submit it for processing and its progress will be monitored in terms of its status (ready, executing, staging, completed) (Figure 5). The ED may also need to monitor the state of processing resources. During processing the should be able to stop or cancel the workflow.

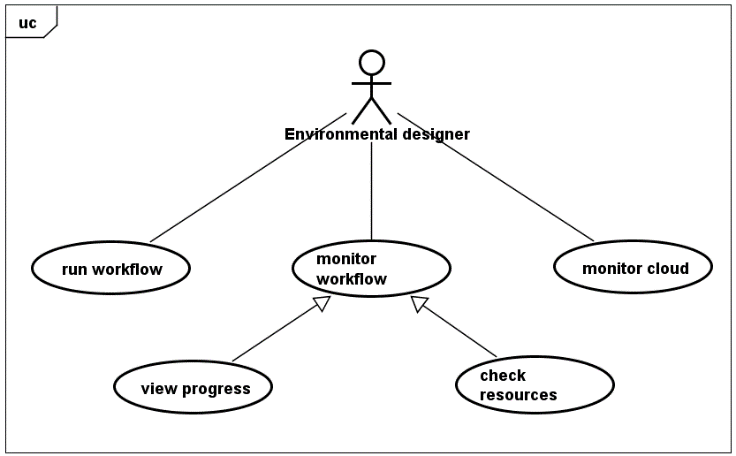


Figure 8 Run and monitor workflow

### Output + visualise results

Following analysis results from the workflow shall be stored and accessible to the ED (Figure 6). Numerical and statistical summaries of the results will be generated by the ED using one of a range of predefined methods. These results would be stored and available for download in formats that can be further analysed or shared in spreadsheets. The ED may need to use simple visualisation of results (histograms, pie charts, line charts). Visualisation tools should include the ability to represent results by geospatial mapping. The graphics generated will be downloaded as high-quality images or vector graphics and used in reports. Within the interface the ED may need to compare two or more visualisations side-by-side to undertake a visual analysis. The ED also needs to share visualisation results with collaborators by providing a secure link to a webpage where the graphics can be viewed and download.

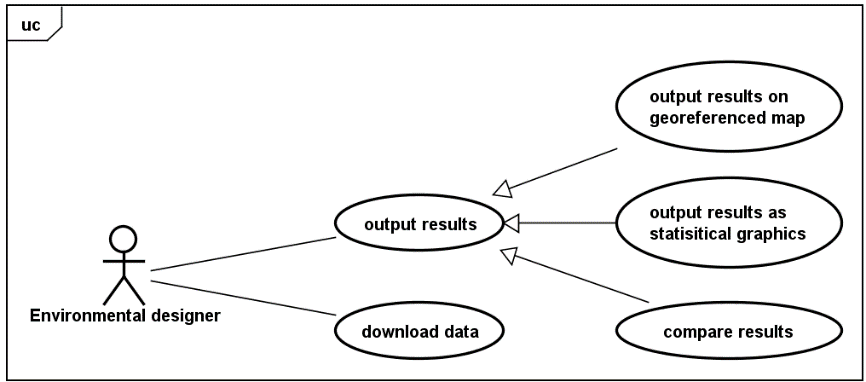


Figure 9 Output and visualise results

### Manage design strategies

The design strategies specific to individual ED’s, it is necessary, therefore, to manage the predefined general design strategies used by the system (Figure 7). Some EDs will wish to add new strategies, removing unwanted strategies and editing existing ones. A standard unambiguous method for describing a strategy is required.

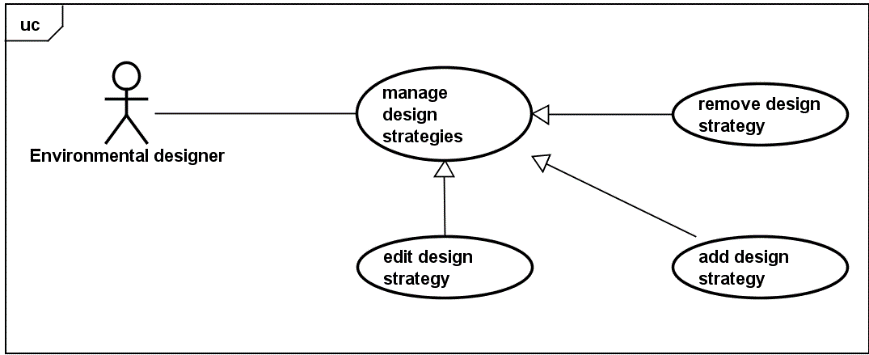


Figure 10 Manage design strategies

### Proposed System Architecture

Figure 8 shows the high-level system architecture based around Infrastructure as a Service (IaaS) provided by AWS cloud infrastructure. The prototype focuses on the items shown in red: workflow and design strategy management system, visualisation and analysis. Each of which communicate with elements of the AWS Cloud via the API, S3 for reading and writing data and EMR for submitting and controlling analytics. AWS Physical hardware on Elastic Compute is indirectly via configurations specified for EMR. Full security for a range of users is considered beyond the scope of the project, however for local prototype application development AWS Credentials are required to access the AWS API. Credentials are stored locally and automatically instantiated using the AWS SDK toolkit within the development environment. Each of the key elements in the overall architecture can be decomposed into lower level subcomponents shown in Figure 9.

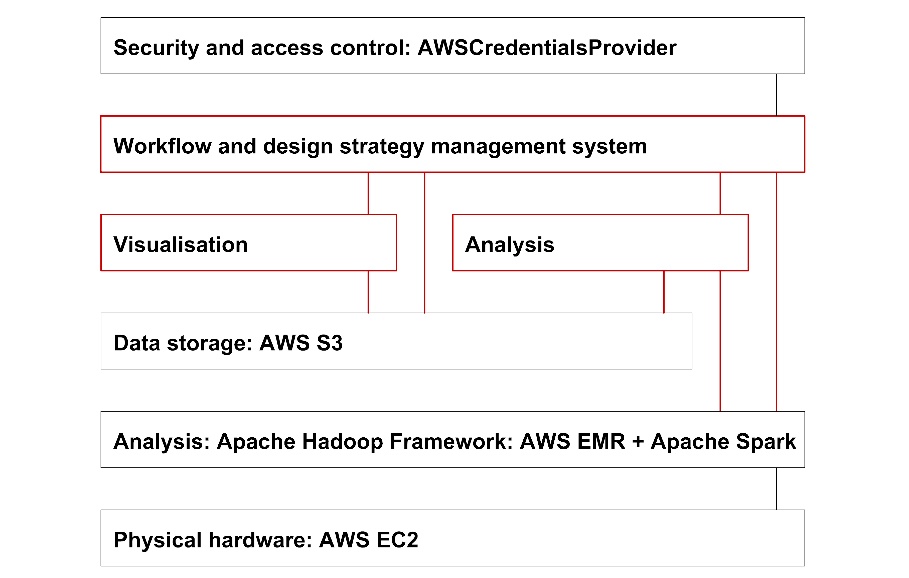


Figure 11 Proposed architecture

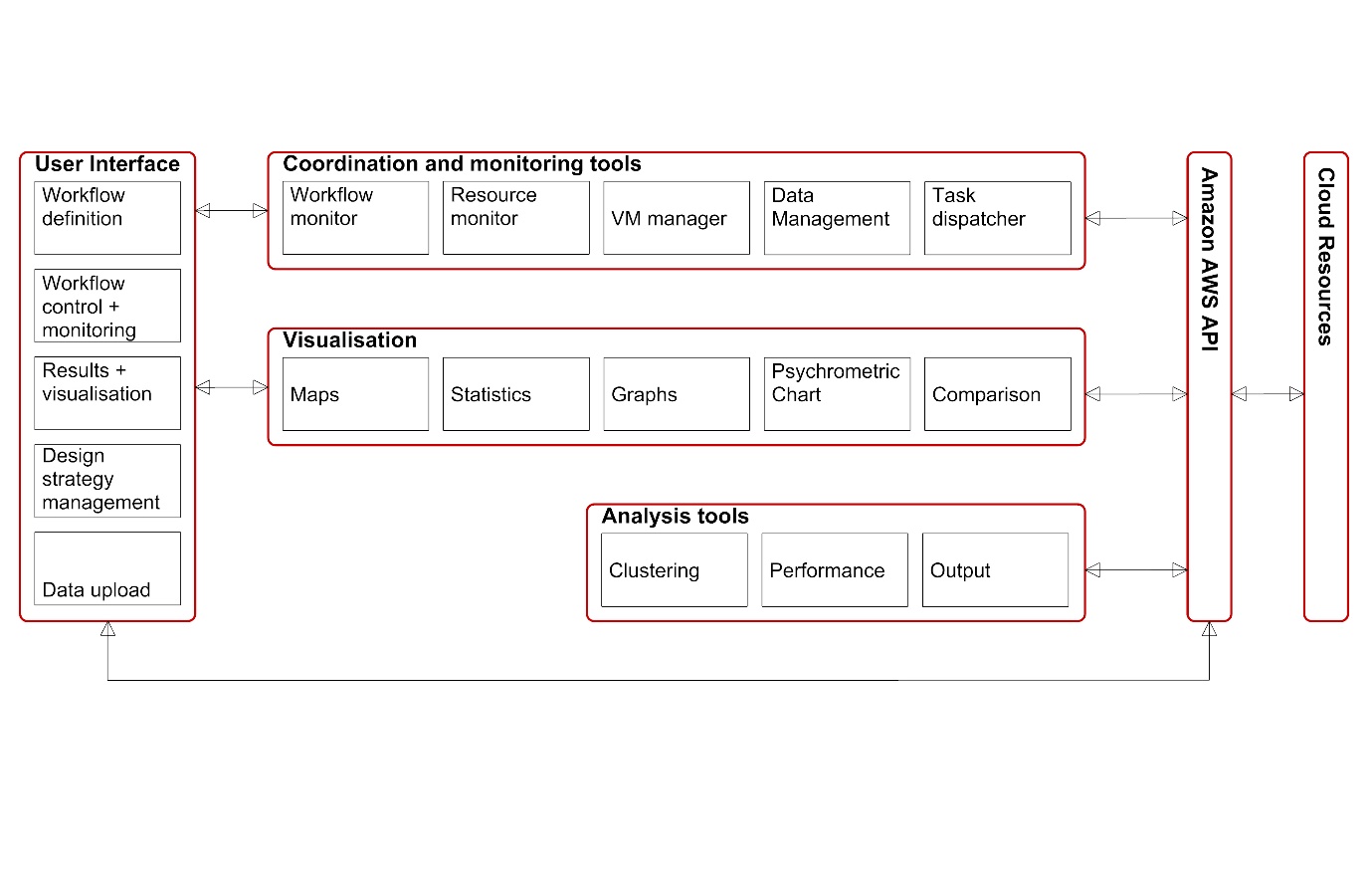


Figure 12 Architecture for the system

Key architectural elements define the proposed system packages (Figure 11). The sub-package within the user interface package, Workflowbuilder, contains domain specific components for climate-driven building design and is regarded as an interchangeable element if the system was applied to another domain.

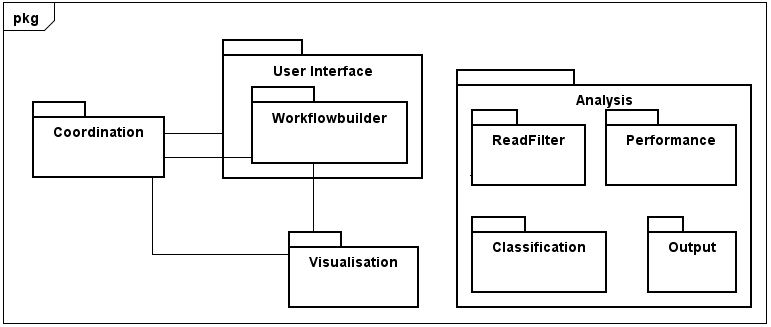


Figure 13 Simplified package diagram.

## Sequence Diagrams for General use cases

The following sequence diagrams (Figure 12, Figure 13, Figure 14, Figure 15) show how the use cases described in (ref to section) have been implemented in the final prototype using the classes and packages described in the previous section.

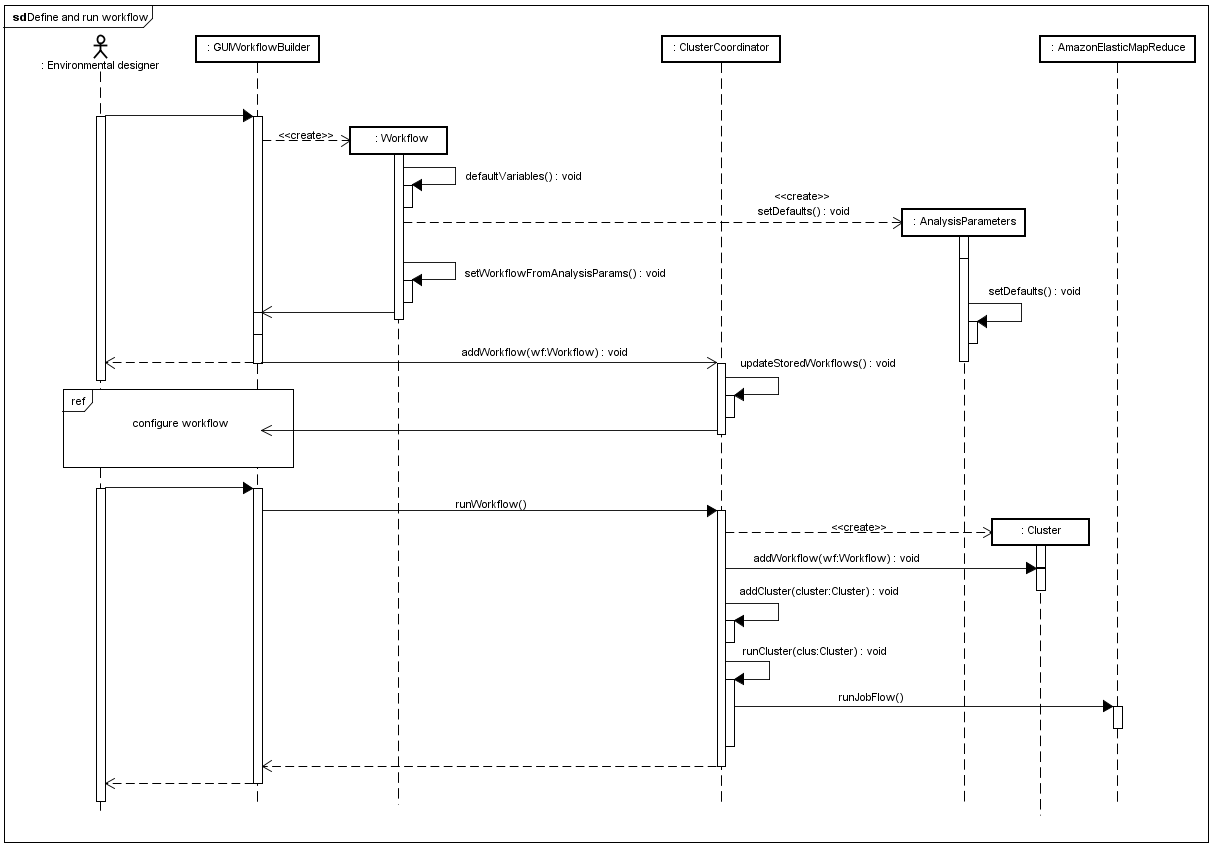


Figure 14 Sequence diagram for defining and running a workflow

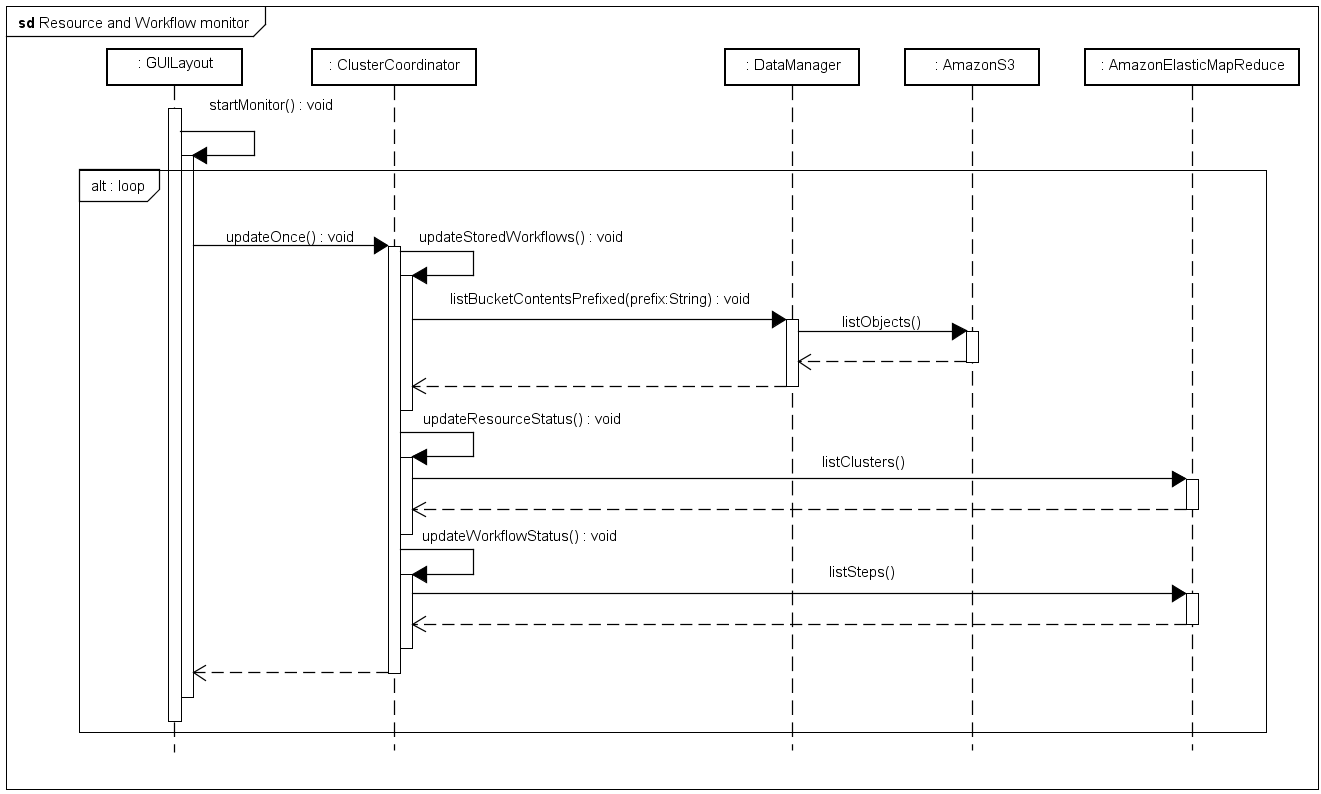


Figure 15 Sequence diagram for monitoring workflows and resources.

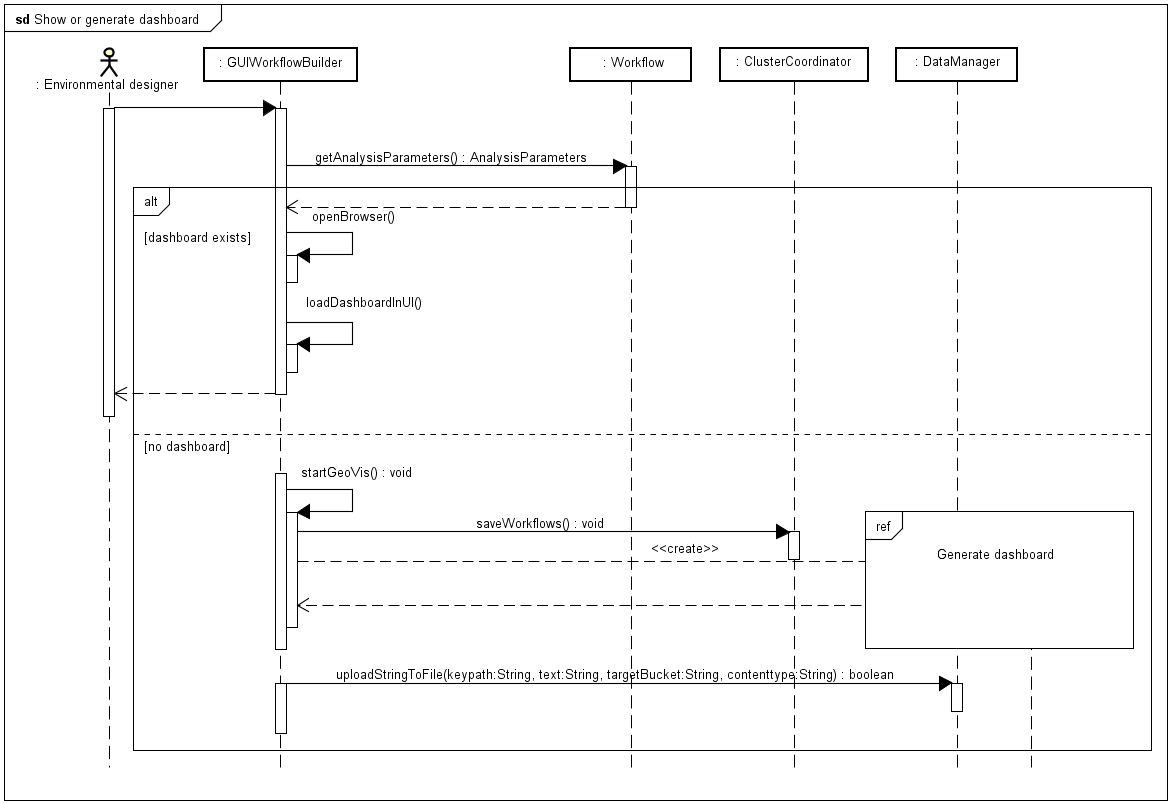


Figure 16 Sequence diagram for show or generate dashboard.

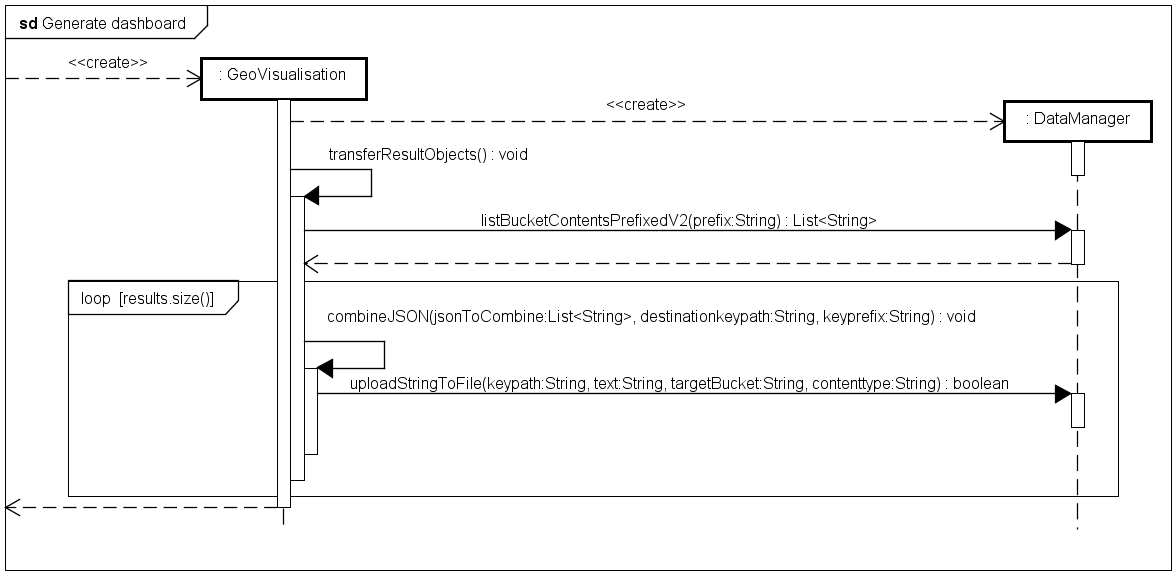


Figure 17 Sequence diagram for generate dashboard also see Figure 14.

## Design methodology

The design process was broken into a series of major development increments (see Figure 16). Each increment lasted approximately 25 days and followed Agile Model Driven Development (AMDD) with UML (Ambler, 2004). First, over several days, high-level modelling was used to understand the scope, requirements and potential architecture of the system. This was followed by a series of construction iterations, each of which began with a planning phase. Requirements were ranked by priority and the highest priority implemented first. Over several hours, UML models were produced to explore what should be built for the iteration and to estimate the time required. Issues identified in the planning models were then developed in more detail using just-in-time models, created in less than thirty minutes involving hand-sketched flow diagrams, sequence diagrams and class diagrams. Using these modelled details code was written during the following hours or days using a test-first and refactor approach.

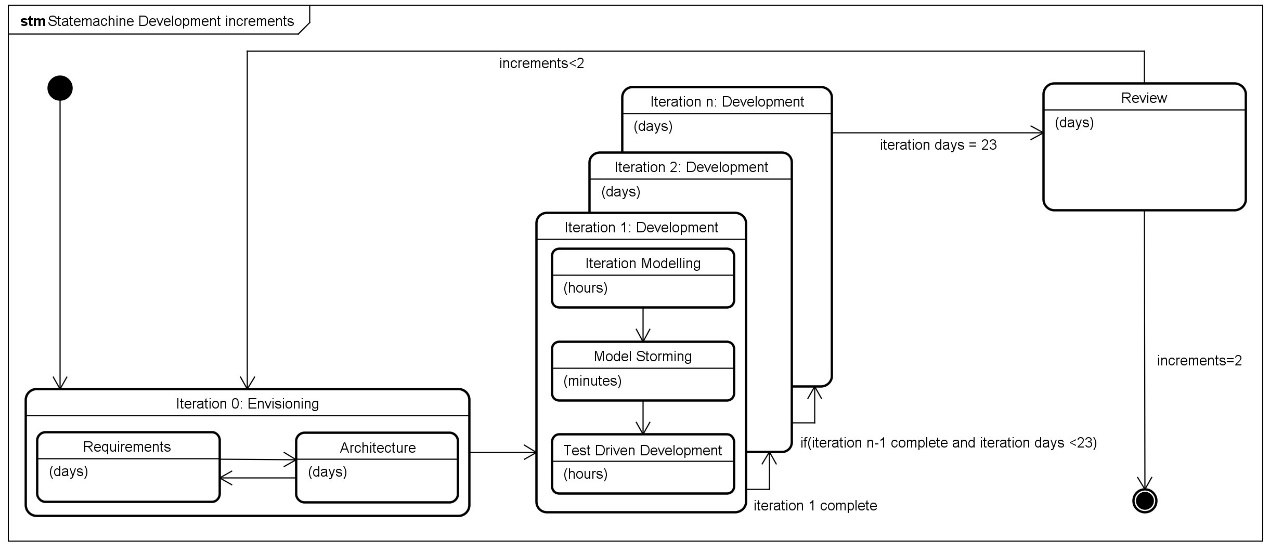


Figure 18 Development increments (extended from Ambler (2004), p119).

# Implementation

## Development iterations

Implementation of the software prototype followed the AMDD methodology over five main iterations each including prioritising requirements, UML modelling, code development, integration testing and refactoring:

1. Key challenges across all packages were addressed, with the goal of a prototype that could perform all basic functionality. Functions were developed to use key aspects of the AWS API, reading and writing data in S3 and programmatically running clusters with simple Spark analysis routines. A rudimentary user interface was defined using Java Swing. This enabled predefined analysis routines and cluster configurations to be submitted to EMR, results written to S3 and visualised on an online map using HTML, JavaScript and D3.js.
2. Development focused on the Workflowbuilder package, monitoring of resources and workflow status within the Coordination Package and structuring the Analysis Package. UI development switched to JavaFX and the interface for user configuration of analysis workflows was developed and tested. An analysis pipeline was developed to provide parsing and filtering of the data set based on analysis parameters associated with workflows. Basic performance metrics were implemented with Apache Spark. Output from analysis was structured in preparation for developing a graphical visualisation framework.
3. Visualisation package was developed including routines to read, reformat and write output from analysis to a public server space. A dashboard framework was developed to read and display output data and provide user interaction.
4. Refactoring the analysis package to provide additional mechanisms to evaluate results.
5. The dashboard framework in the visualisation package was restructured with more code reuse, allowing multiple controllable instances of graphic tools simultaneously. Improving the graphical representation and user access to the evaluation metrics generated through analysis.

## Environment

The prototype operates locally on a Java Virtual Machine (JVM), analytics take place within AWS Cloud using the EMR framework running spark-core\_2.30. Hardware on EMR is configured by the user via the local application as part of the workflow definition, users have access to all classes of AWS EMR instances and can run up to 20 instances (1 master and 19 slaves).

The system was developed and tested on Windows 10 Pro running on 64-bit Operating System with Intel Core i7-8700k CPU and 16Gb ram. Java development took place using Eclipse photon IDE with AWS Toolkit for Eclipse (*AWS Toolkit for Eclipse*, no date). UML models were developed using Astah Community with some additional diagrams modelled in Rhino3d. JavaScript, html and CSS were used to create the dashboard web framework that was developed using Sublime Text 3 and tested with http-server (*http-server: a command-line http server*, 2018) to debug locally within Chrome 69. Github.com was used for version control and the latest version of code and documentation is available: <https://github.com/rolyhudson/climacolombia.git>.

## Architecture

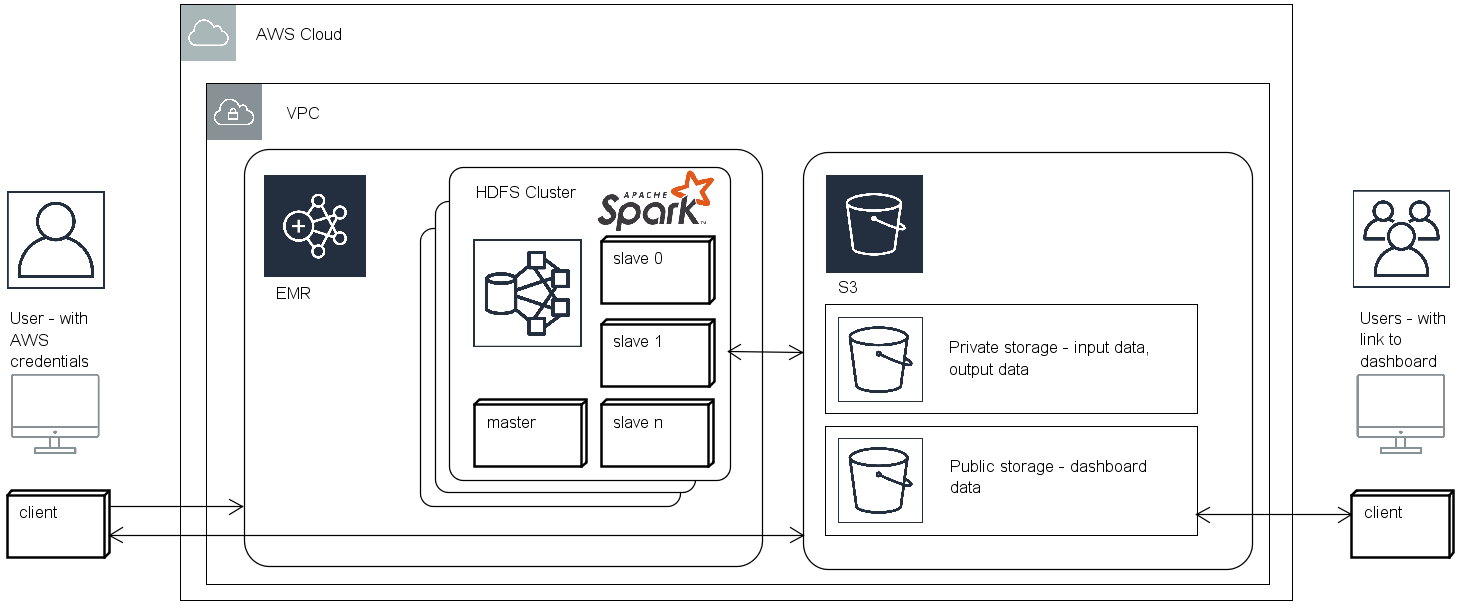


Figure 19 Deployment on AWS

The application uses IaaS on AWS cloud infrastructure. Users access the local application (with AWS login credentials) and can launch resources within Amazon’s VPC. A graphical interface enables users to create workflows (analytic jobs) and configure the hardware that will be used to compute them. These analytic jobs are launched as Hadoop distributed file system (HDFS) clusters on Amazons’ EMR each running Apache Spark, multiple clusters be can be configured and launched simultaneously. Clusters have access input data and store results from the analytic jobs in private buckets within AWS S3. When analytics on a cluster completes, output files are transformed and stored in a publicly accessible bucket with a web framework that defines a graphical dashboard for visualising results. A Uniform Resource Locator ([URL](https://en.wikipedia.org/wiki/URL)) is created for the dashboard that can be shared by the original user so others can access the visualisation and results.

## Packages

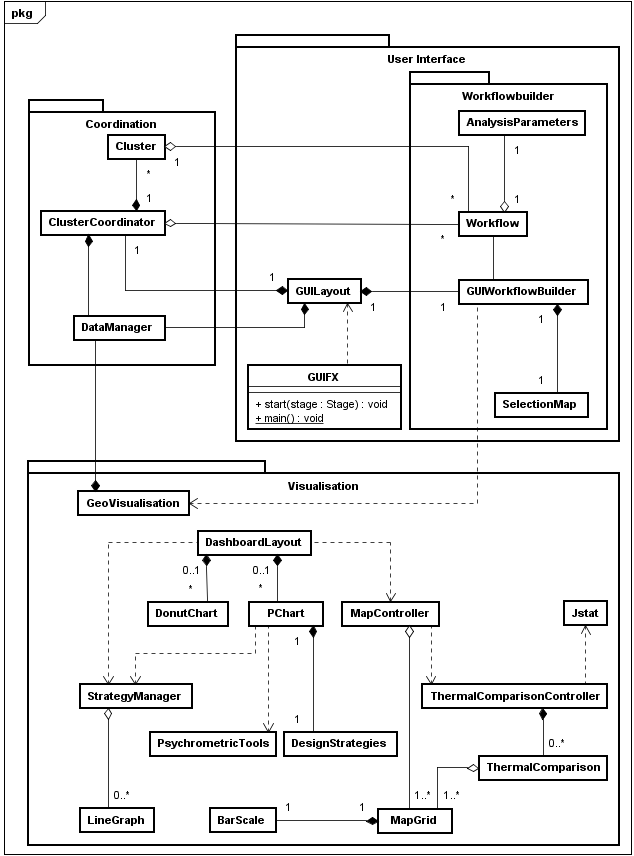


Figure 20 Coordination, User Interface and Visualisation packages overview

The system is implemented with four key packages (Figure 19 and Figure 20). The User Interface package provides functionality that presents the status of cloud resources and running analysis jobs. New analysis routines and data sets can be uploaded from the UI as .jar packages. The UI also provides access to dashboard of results via an embedded web engine and representation (and a mocked-up editor) of the design strategies editor.

The main component in the User Interface package is a sub-package, Workflowbuilder, which concerns all classes involved with the definition and of a workflow which include defining spatial zones for analysis and specifying a series of parameters concerning the data, analysis method and temporal scales. The coordination package interfaces with the EMR client via the ClusterCoordinator class and S3 via the DataManager class both of which use locally stored credentials and the AWSCredentialManager for secure access to AWS services.

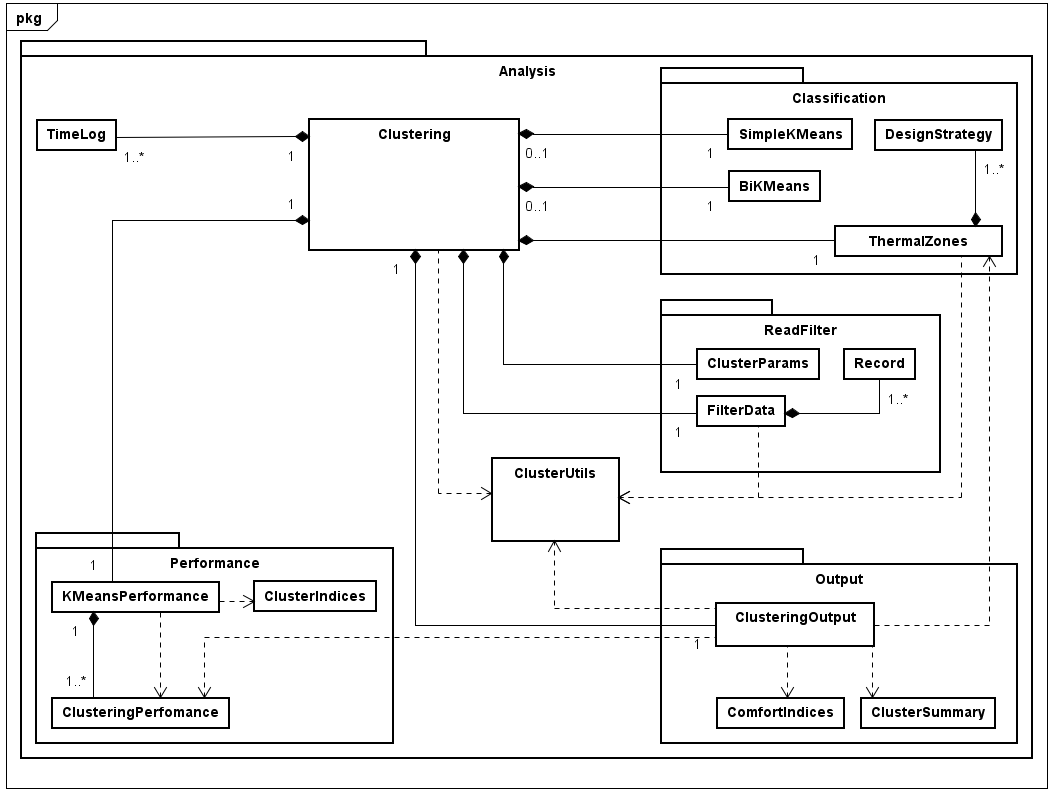


Figure 21 Analysis package overview

The Analysis package is independent of other packages and defines the analytical procedure used by Apache Spark running on AWS EMR. It includes four sub-packages each handling different aspects of the analytic routine. Analysis parameters defined by the User Interface are parsed and filtered by classes within the Read Filter sub-package. The Performance sub-package computes the performance of clustering solution using a series of standard indices, additionally the number of clusters can be optimised base on one of these indices. Classification sub-package implements two clustering methods from Spark’s Machine Learning library MLLib (*MLlib: Main Guide - Spark 2.3.2 Documentation*, no date) K-Means clustering and Bisecting K-means clustering and a third method that is a hybrid of the previous two. Data vectors are submitted to the Classification sub-package, clustering models are constructed, and the data is classified and assigned applicable design strategies. Finally, the Output sub-package organises the results and the meta-data related to the performance of the process and writes these in a logical storage structure to S3.

Once analysis completes the GeoVisualisation class in the Visualisation package is used to post- process analytic output and move files to public buckets on S3 with a unique URL that is provided to the user. A webpage is created which references the dashboard framework components (also hosted within a public area of S3) and provides an interactive, online interface for visualising the results.

## Climate data

The data set is based on multivariate, historical, monthly averages of gridded climate data from three sources. First, climatic data from 1901-2009 formatted as ESRI ASCII raster by CGIAR CSI (Cgiar-csi.org., 2012) based on original data from CRU (Jones and Harris, 2008). Secondly, wind speeds from the CCMP gridded surface vector winds (Wentz *et al.*, 2015). Thirdly, elevation data is extracted from a hole-filled DEM of SRTM (Jarvis, A., H.I. Reuter, A. Nelson, 2008). Figure 18 illustrates the data preparation steps. A C# dot net program was written that takes a topojson (Bostock, 2017b) format file as input, this describes the boundary (or collection of boundaries) that define the zone of interest. A point grid is generated at half degree latitude and longitude intervals filling the area(s) of study. Cross-referencing the grid to the DEM determines altitudes for each point.

The grid is used to extract climate data from a selected period from the CRU and CCMP datasets. CRU data is formatted as ASCII ESRI raster format at the same resolution as the grid. CCMP data is in netCDF format, this was pre-processed with an independent Java program, written using the Unidata (2012) netCFD Java library. The netCDF data was converted to the ASCII ESRI raster format. Each ASCII raster file represents a single month of a year and contains data for earth’s surface. Data points in the raster files that coincide with our grid vertices are found and stored with the georeferenced grid in arrays. Relative humidity derived using the ratio between vapour pressure and saturation pressure. The prepared climate data is saved as a Comma Separated Value (csv) file.

Source code for the data preparation is available at: <https://github.com/rolyhudson/griddedClimateDataProcessing.git>

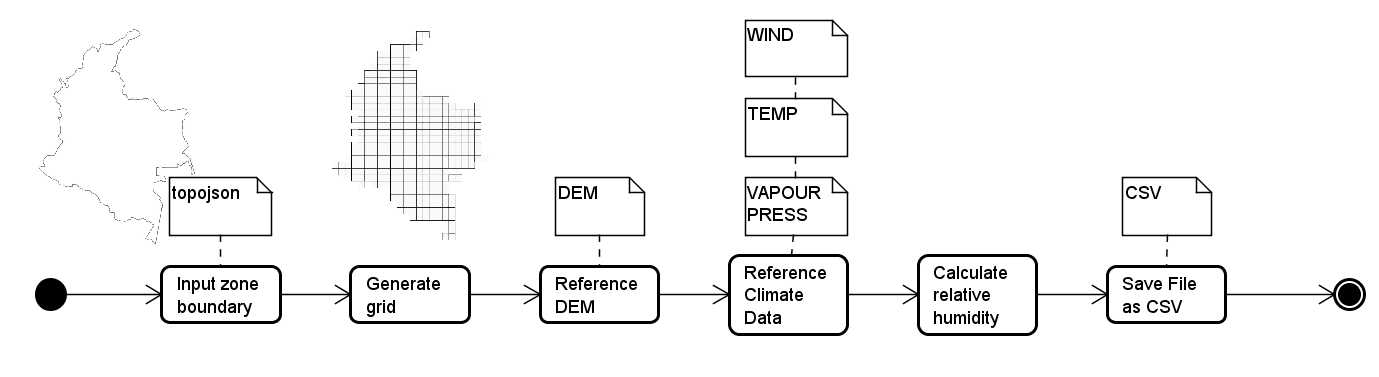


Figure 22 Flow diagram showing the stages of the data preparation process.

## Workflow management system

### Workflowbuilder and Coordination package class diagrams

At the core of the application is the workflow management system consisting of classes defined with the Workflowbuilder package (Figure 21 and Figure 22) which communicate with the classes in the Coordination Package (Figure 23).

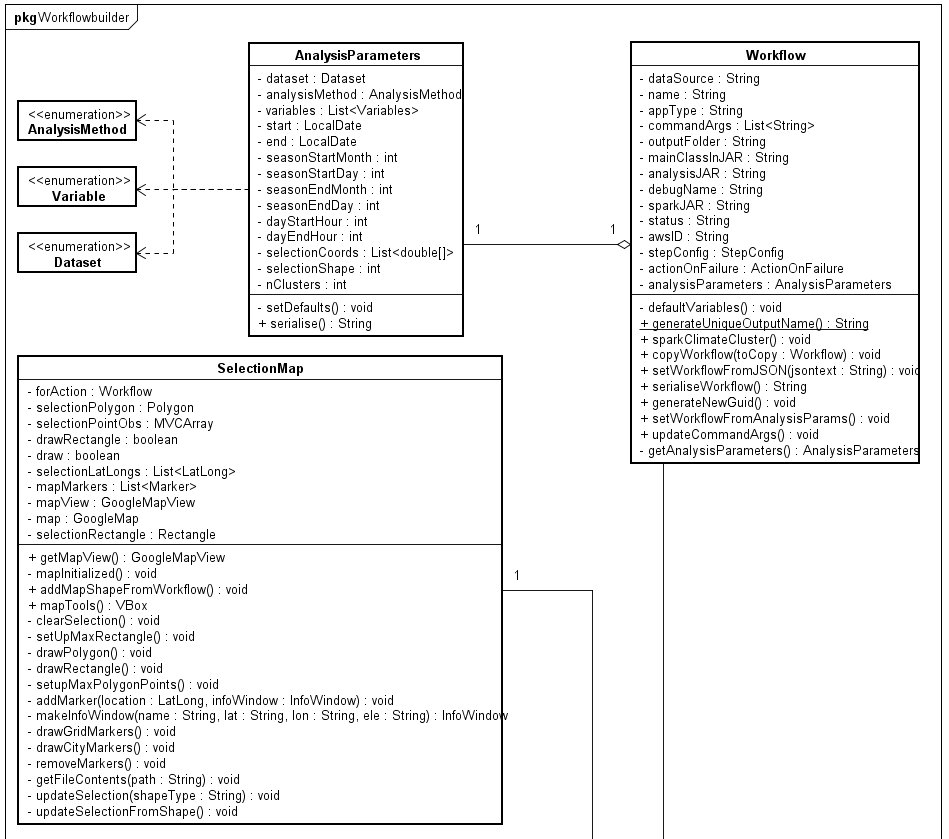


Figure 23 Workflowbuilder package class diagrams a



Figure 24 Workflowbuilder package class diagrams b

### Workflow creation

The Workflow class manages all the parameters for an AWS Step, additionally, the Workflow class includes a single instance of the AnalysisParameters as the attribute analysisParameters. The AnalysisParameters is a description of the domain specific configuration of the workflow. The user can configure this class instance via the GUI (Figure 24) which provides access to hardware configuration options, choice of data set, selection of variables for clustering, clustering method, number of clusters and control of the temporal range of the analysis. Spatial specification of the analytics is via the map an instance of the SelectionMap class that implements GoogleMaps in JavaFX using GMapsFx (Terpilowski, no date). The map shows the selected data set, locations as red and a selection polygon or rectangle can be drawn and edited to define the region for analysis.

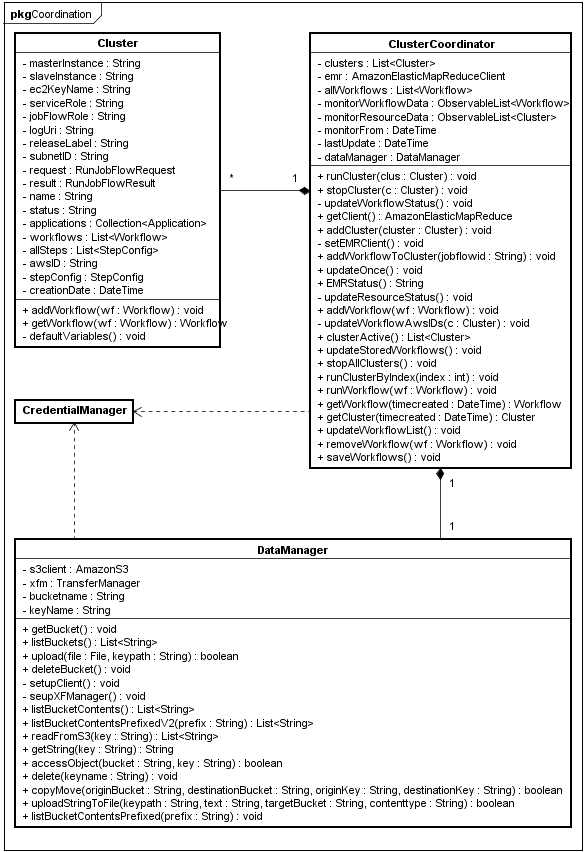


Figure 25 Coordination package class diagrams

User edits to the workflow via the UI are stored in a list of workflows maintained by the ClusterCoordinator class. Multiple workflows can be created and saved in one session. On exiting the application session all workflows are stored in S3 in a private bucket in JSON format. Each workflow instance has a GUID which determines the S3 object name, the object contains a string in JSON format that is returned by the class’ serialise method. Workflow objects are pushed to S3 by the UI’s instance of the DataManager class datamanager.

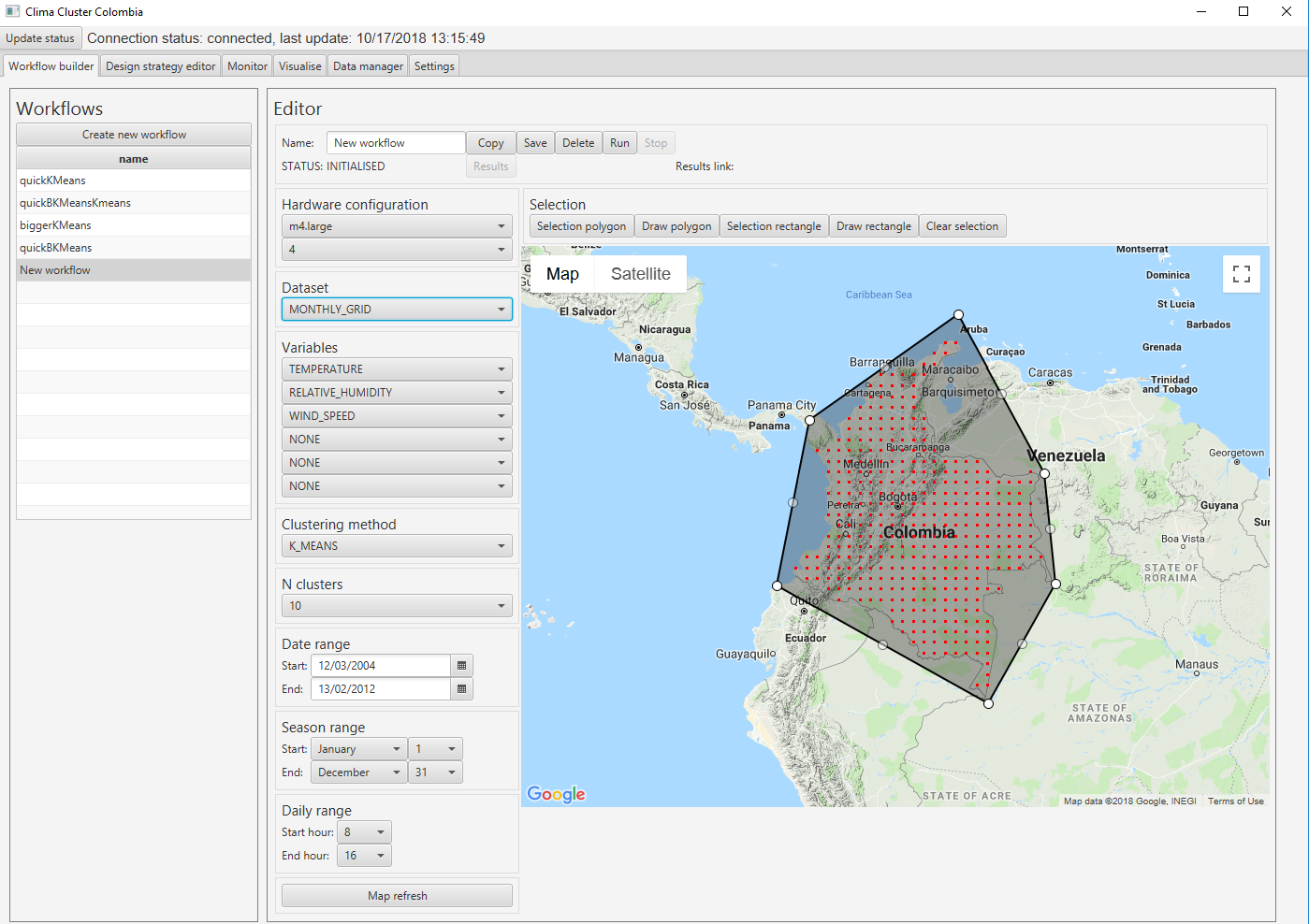


Figure 26 GUI for Workflowbuilder

### Monitoring resources and workflows

The ClusterCoordinator maintains a list of all EMR clusters created after a datetime parameter – monitorFrom (set by the user). Each EMR cluster is represented locally by an instance of the Cluster class. This manages all the parameters required to create a new EMR cluster (Figure 23). While the application is running the ClusterCoordinator is constantly monitoring the AWS resources (see Figure 13). The goal of the monitoring process is to keep the ClusterCoordinator’s locally stored lists of clusters and workflows up to date with the status of the EMR Clusters and associated steps found on the cloud. The implemented monitoring algorithm (Figure 25) allows users to break between work sessions and shut down the application, on returning the status of any running clusters or workflows is reported and the user can continue to edit, copy or execute a workflow created in a previous session.

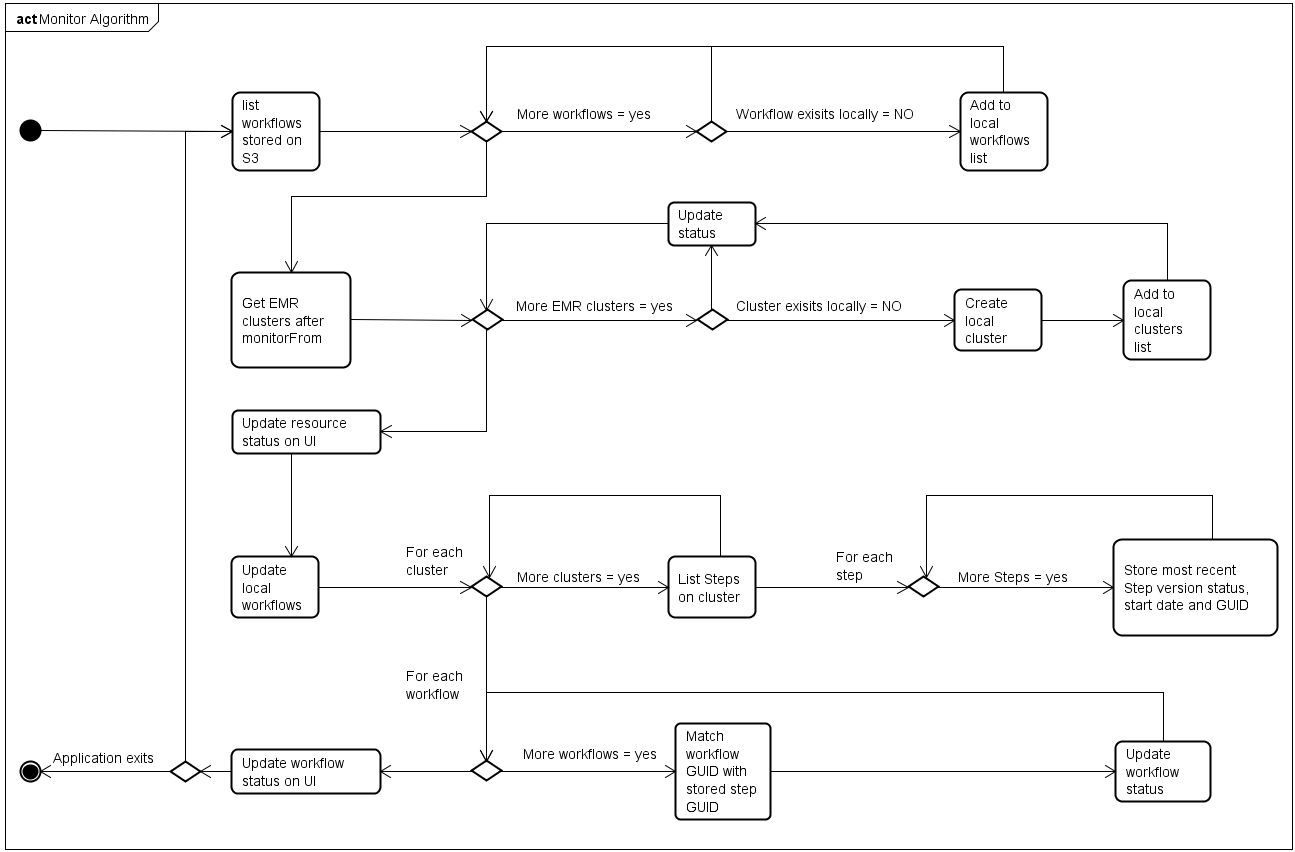


Figure 27 Resource and workflow monitoring algorithm

Resource monitoring takes place on a separate thread to the UI and is executed every minute to avoid AWS’s throttling limits. Key to this process is the workflow class GUID and the method, setWorkflowFromJSON, which creates a new workflow by de-serialising the JSON string stored in the S3 object. Once locally stored lists are updated they are added to an ObservableList that is used to populate the UI’s monitor tab (Figure 26). To edit a workflow, it must be selected from the left-hand panel in the Workflow builder tab (Figure 24). The status of a workflow is shown in the UI, its state (Figure 27) determines the availability of controls on the UI. For example, a COMPLETED workflow only allows the user to copy, delete or generate the dashboard (Figure 28). Whereas an INITALISED workflow provides access to all workflow configuration controls but not the buttons to stop or generate the dashboard. The activity diagram in Figure 29 shows the different processing paths a user can follow when working with a workflow.

### Running a workflow

Once the user has prepared their workflow they click the run button on the UI and the workflow is sent to EMR by the ClusterCoordinator. The workflow can either be added to an existing cluster or a new cluster can be created if none exist or the user requires different hardware configuration. Figure 29 illustrates the steps involved in running a workflow.

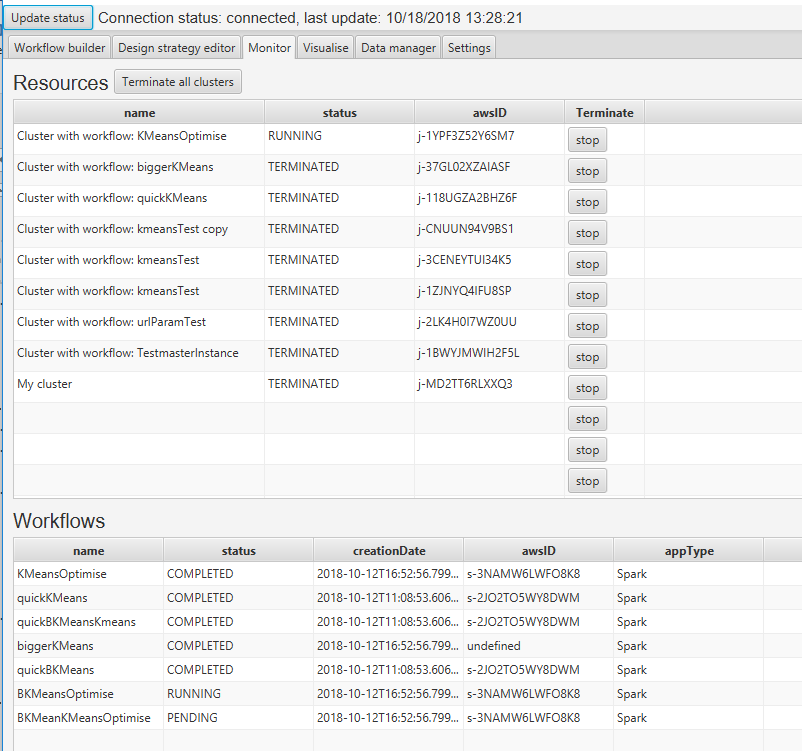


Figure 28 UI Monitor resources and workflows tab

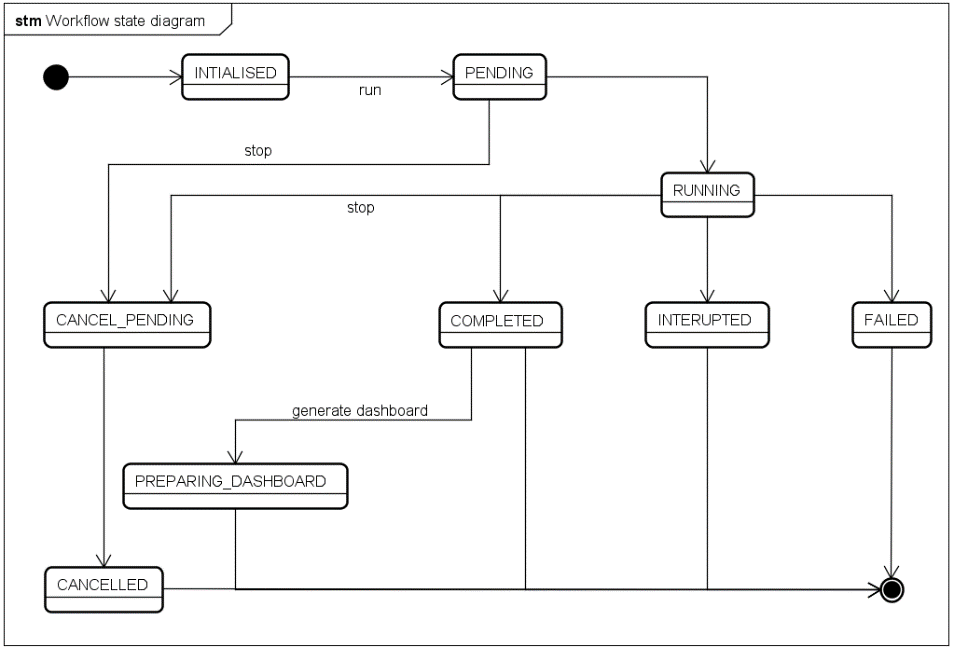


Figure 29 Workflow state machine

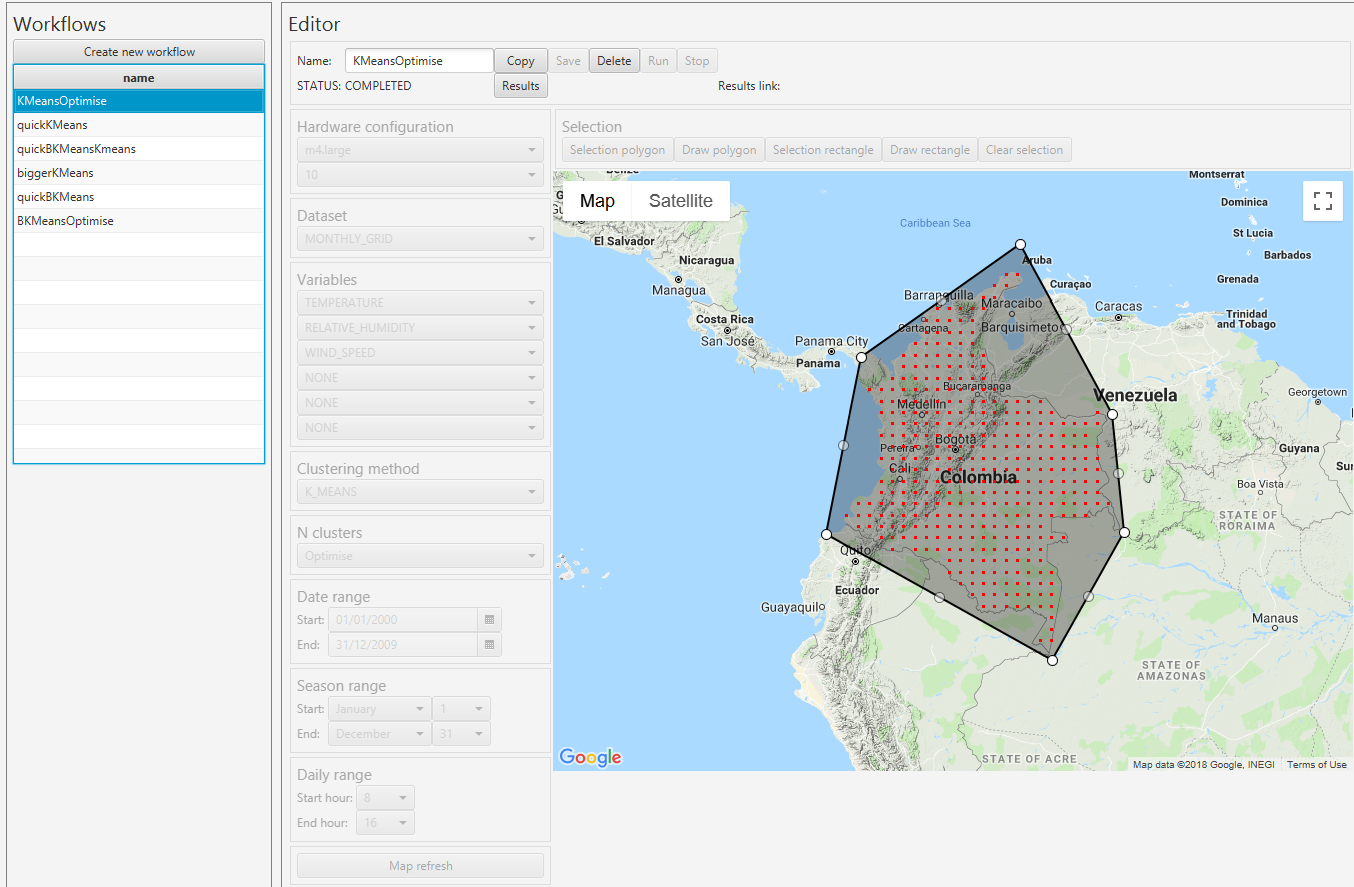


Figure 30 Workflow with status COMPLETED

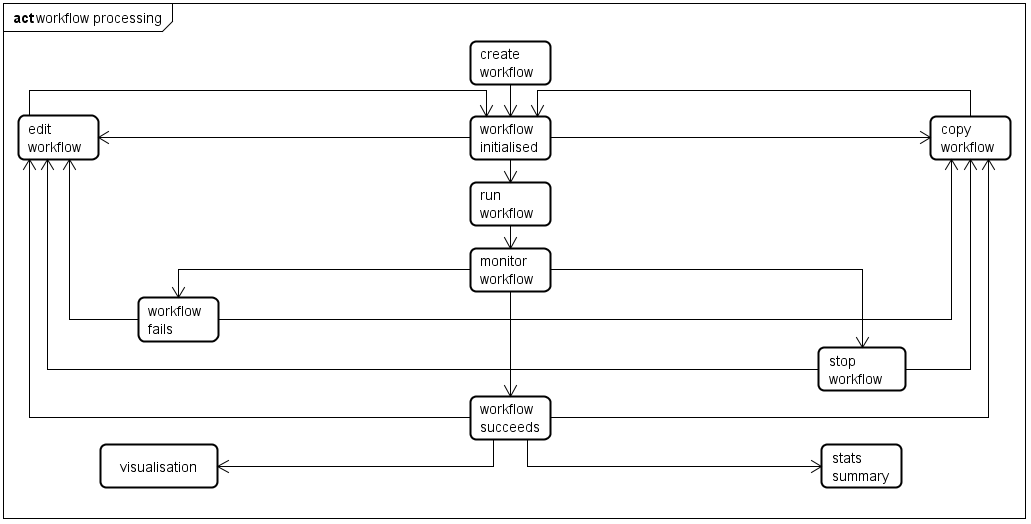


Figure 31 Processing workflows activity diagram

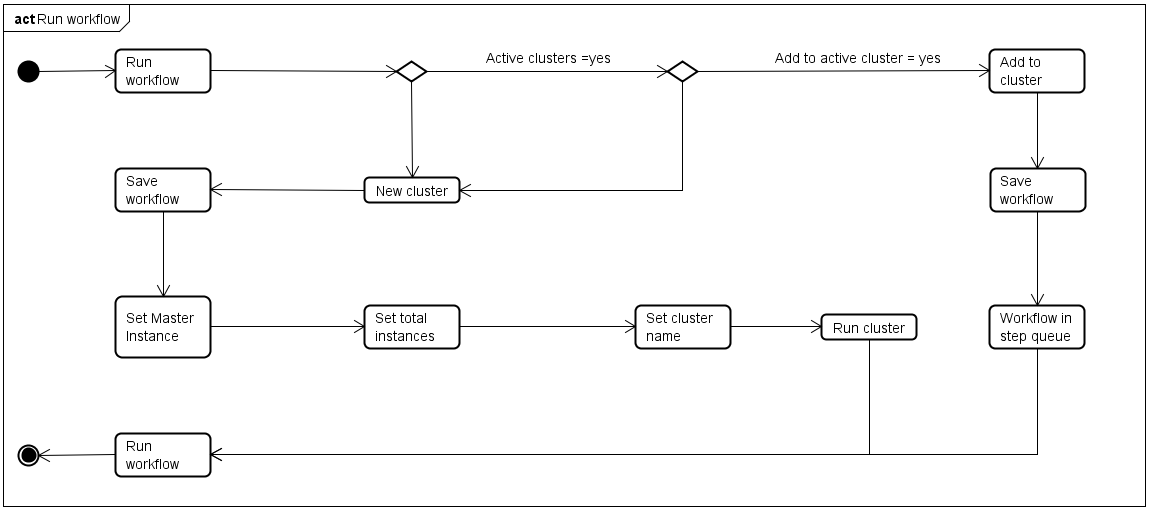


Figure 32 Run workflow activity diagram

### Management of design strategies

The management of design strategies has not been fully developed for the prototype. The UI includes a tab for editing strategies, however, this is a mock-up of a simple drawing interface that would allow users to draw, edit, import and save strategies over a psychrometric chart (Figure 31). The intention is that each strategy is represented as closed polygon with a name, an array of these polygons can be stored, one per line, in a simple csv file where the first field is the name followed by x and y coordinates of the vertices. For the purposes of the prototype the strategies described by Manzano-Agugliaro *et al.*, (2015) have been stored in this format and are used for the analytic process described in the following section.

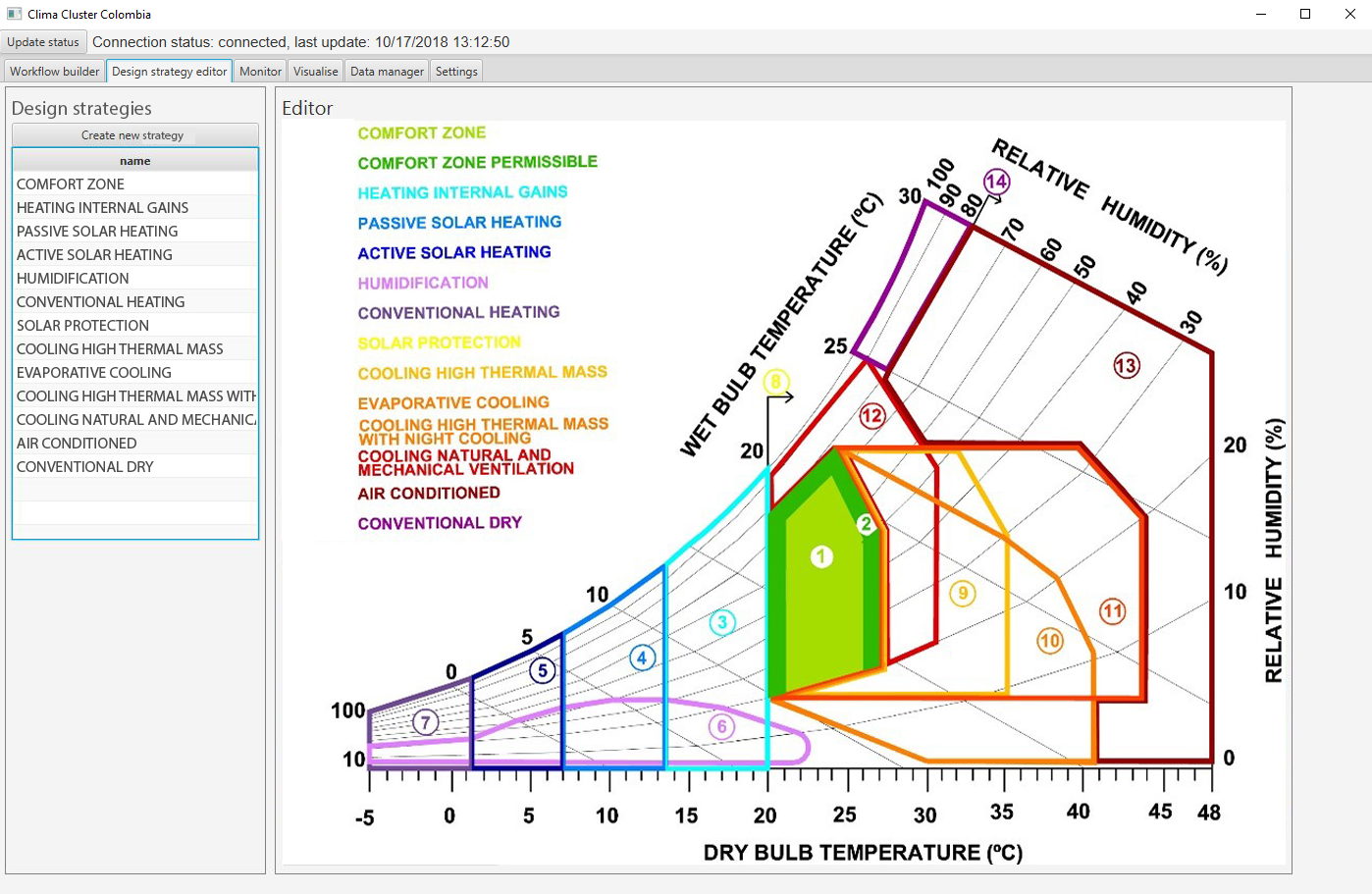


Figure 33 Design strategy editor mock-up.

## Analytics

The analytic pipeline is a Spark application defined as an independent Java project [[1]](#footnote-1) which is exported locally as a \*.jar and uploaded to S3 via the UI with the keypath “sparkJAR”. When a workflow is submitted to EMR it defines parameters for the configuration of an EMR Step. These parameters include its name, the action that the cluster takes on failure, the name of the analysis \*.jar and a set of arguments. The analysis \*.jar parameter refers to Amazon’s command-runner.jar that enables spark-submit script. The spark-submit script is defined in the arguments that specify:

1. the deployment mode
2. the class in the application that contains the main method
3. the location of the spark application jar
4. Data source
5. Output folder
6. Location of the workflow file on S3
7. Location of the design strategy file on S3

These seven arguments can be seen in the EMR management console within the cluster details (Figure 32). Once the EMR cluster is running arguments 4-7 are passed to the class in the application that contains the main method as an array of Strings. The Clustering class (Figure 33) includes the main method and acts as the controller for the analytics process, it initialises the SparkSession and parses the clustering parameters, reads and filters the data set and the design strategies. Depending on the selected clustering method, performance is evaluated before the selected data is classified and design strategies appropriate to each cluster and data-point are identified and summarised. Finally, the results are written to S3. Figure 22 shows the class structure of the Analysis package, Figure 36 illustrates the sequence of processes used to implement the analysis.

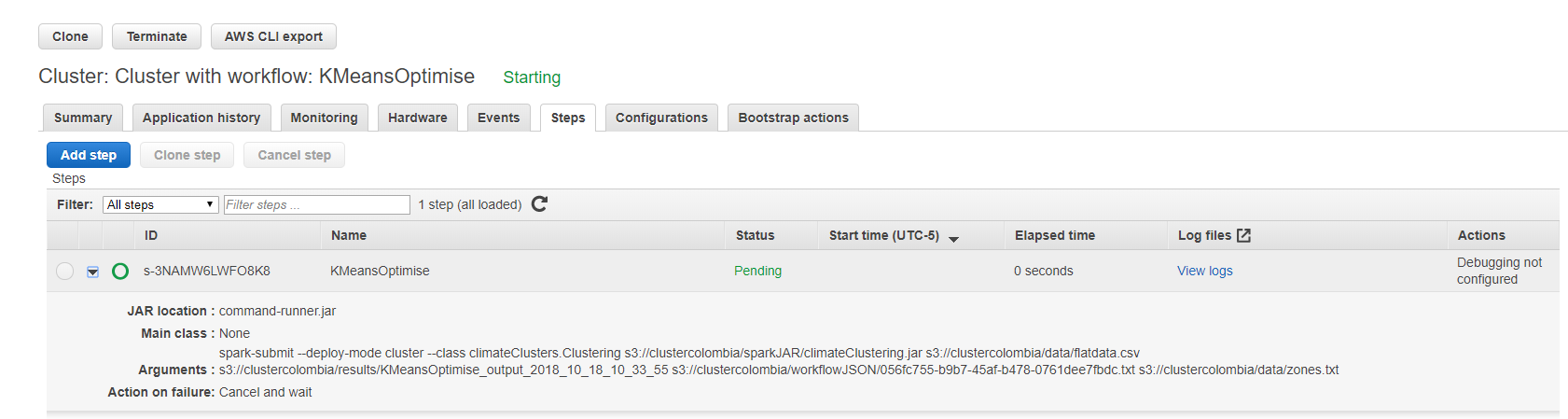


Figure 34 Parameters for a Step on EMR

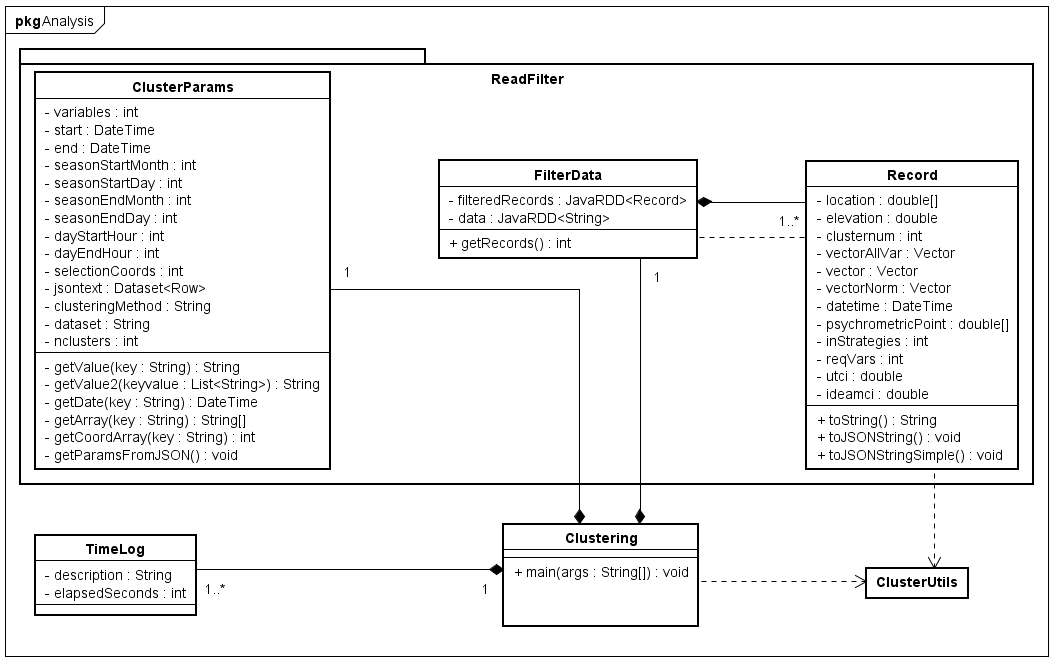


Figure 35 Analysis package and ReadFilter sub-package.

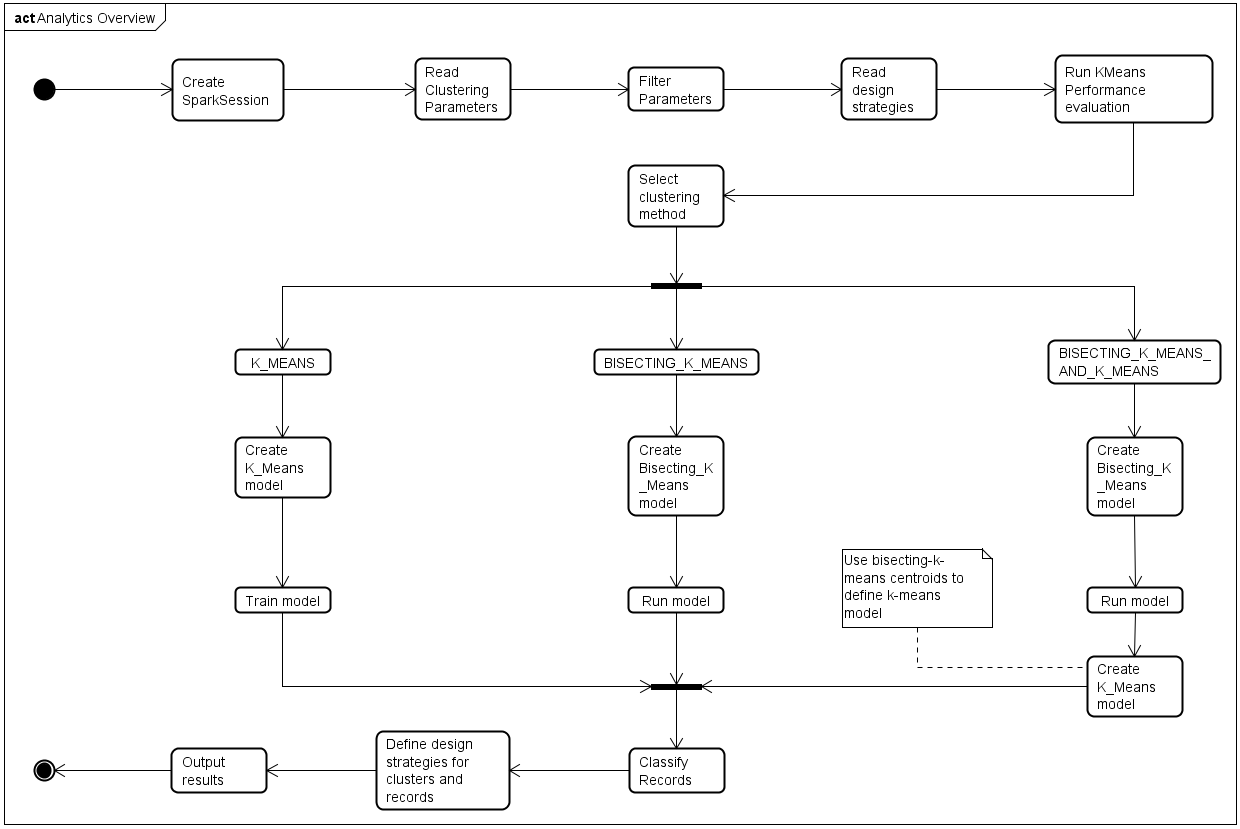


Figure 36 Analytics overview activity diagram

### Clustering Parameters and Filter Data

Deserialization of the Workflow class and its analysisParameters attribute is undertaken by the ClusterParams class (Figure 33) by reading from a string in JSON format, stored on S3. The JSON file is read directly to a Dataset<Row> and from that the clustering method, number of clusters, temporal and spatial configuration of the workflow can be selected.

The FilterData class (Figure 33) takes the parsed clustering parameters, reads the input file containing all the climate data in csv format. Using a combination of the RDD operations; filter and map, and the helper methods defined in the ClusterUtils class a subset of the original data is defined as a JavaRDD of Record objects. This dataset contains only data points within the required spatial and temporal ranges.

The Record class (Figure 33) includes attributes to store all the variables contained on each line of the required input data and identifies the variables to be used for clustering. In addition, the Record object stores a psychrometricPoint, a 2d array defined by temperature and relative humidity, a list of design strategies that can be applied to the Record and the comfort indices (UTCI and IDEAMCI) for the record. The Record class includes a method that returns a normalised Vector of the variables destined for the clustering processes (features are normalised with Spark’s MLlib L2 norm). Design strategies stored as a text file on S3 are read and parsed by the ThermalZones class (Figure 37) and stored as a list.

### Clustering Performance Indices

The KMeansPerformance (Figure 35) class is responsible for defining performance indices for the selected clustering method. A workflow can be configured with a user-defined number of clusters or set to optimise the number of clusters. Performance metrics for each solution are stored in a list of ClusteringPerformance objects (Figure 35), an attribute of the KMeansPerformance class. The list either contains results for a single clustering solution or if optimisation has been selected it stores clustering indices for 58 solutions with two to sixty clusters. ClusteringPerformance objects store cost as WSSSE plus Silhouette and Dunn indices. These metrics are calculated by the ClusterIndices class (Figure 35), which includes different methods for each index depending on the selected clustering method.

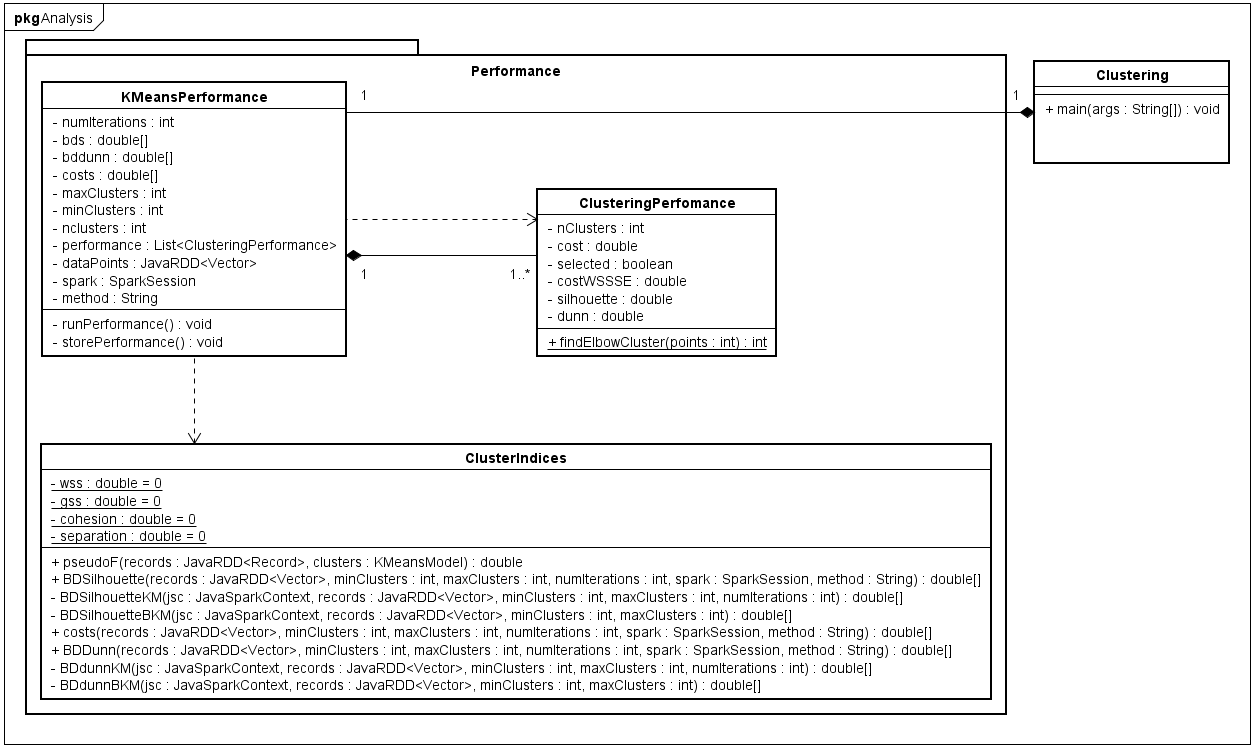


Figure 37 Analysis package and Performance sub-package

Optimisation is based on the WSSSE for each solution and the optimal number of clusters is determined computationally using the elbow method implemented as the static method findElbowCluster in the ClusteringPerformance class. The method uses a geometric algorithm that creates a graph for all solutions and finds the curve inflexion point (clustering solution) furthest from a line subtended between the first and last points. Figure 38 shows a scenario where twelve clusters is found to be the most optimal solution with a WSSSE of 3.7.



Figure 38 Elbow method

### Clustering classification and design strategy assignment

Once performance metrics have been generated, and the optimal number of cluster defined, the analytics proceeds to defining the chosen clustering model. This is based on two classes SimpleKMeans and BiKmeans (Figure 37) which are used independently or combined to provide a third, hybrid method. Each Record is classified (assigned a cluster number) using the selected clustering model and the ThermalZones class tests the Record’s psychrometricPoint to find applicable design strategies and assigns these to the Record. The DesignStrategy class (Figure 37) stores the boundary of a strategy as a list of vertex coordinates defined in terms of temperature and relative humidity. To test if a design strategy is applicable to a Record its psychrometicPoint (a 2d point defined by temperature and relative humidity) can be containment tested against the strategey’s boundary vertices, if the point is inside the polygon the DesignStrategy is added to the Records list.

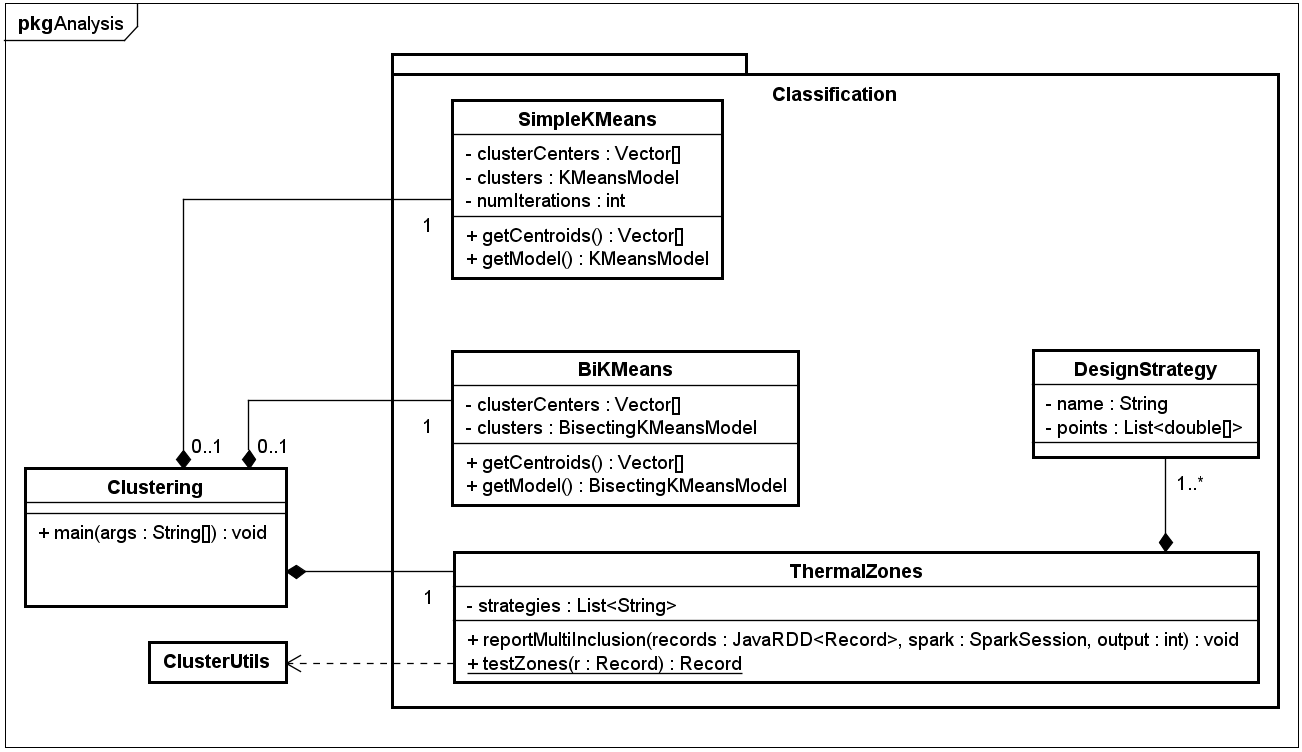


Figure 39 Analysis package and Classification sub-package

### Output of results

In anticipation of querying and quickly visualising the results according to different spatial and temporal ranges, Spark is used to pre-process and structure the output. Meta-data related to the clustering performance and configuration of each cluster is stored. Three key spatiotemporal queries are anticipated. The first, averaging all temporal results to provide a single dataset with no temporal dimensions creating a single typical period for the original temporal scope across the entire spatial range. The next query type creates twelve datasets each representing a typical month, creating a single typical period but with monthly granularity across the entire spatial range. The third anticipated query type provides the finest granularity generating datasets for each year and month and summary information per individual cluster. Lastly, to accommodate users that wish to create custom queries or create their own visualisation processes the full set of results is stored as a single document.

#### Meta-data

Performance metrics stored as a list of ClusteringPerformance objects in the KMeansPerformance class are converted to a Dataset<Row> and then written to a file in JSON format. Using the getComfortIndicesClusters method in the ComfortIndices class (Figure 38), average temperature, relative humidity, thermal indices are defined for each cluster. The comfort indices and the centroids of each cluster are passed to the reportClusterSummary method in the ClusterSummary class (Figure 38). A list of ClusterSummary objects, one for each cluster in the solution, is generated, describing centroids, the number of data points contained, thermal indices, average temperature and average relative humidity. The ClusterSummary object also contains design strategies associated with the cluster. Strategies are determined by using the cluster’s temperature and relative humidity to define a single point and containment testing against the list of DesignStrategy objects in the ThermalZones class. The list of ClusterSummary objects are converted to a Dataset<Row> and then written to a file in JSON format.

#### Typical period

Generating a typical period involves grouping the data first by geographical location. Then for each location finding the most commonly occurring cluster number, that clusters average temperature and relative humidity used to define which design strategies are applicable.

#### Typical period with monthly granularity

Creating a typical period with monthly granularity first requires iterating the months of the year and filtering the data by each month. A typical period for each month is created by grouping by location and then finding the most frequent cluster number and assigning appropriate strategies and storing the results.

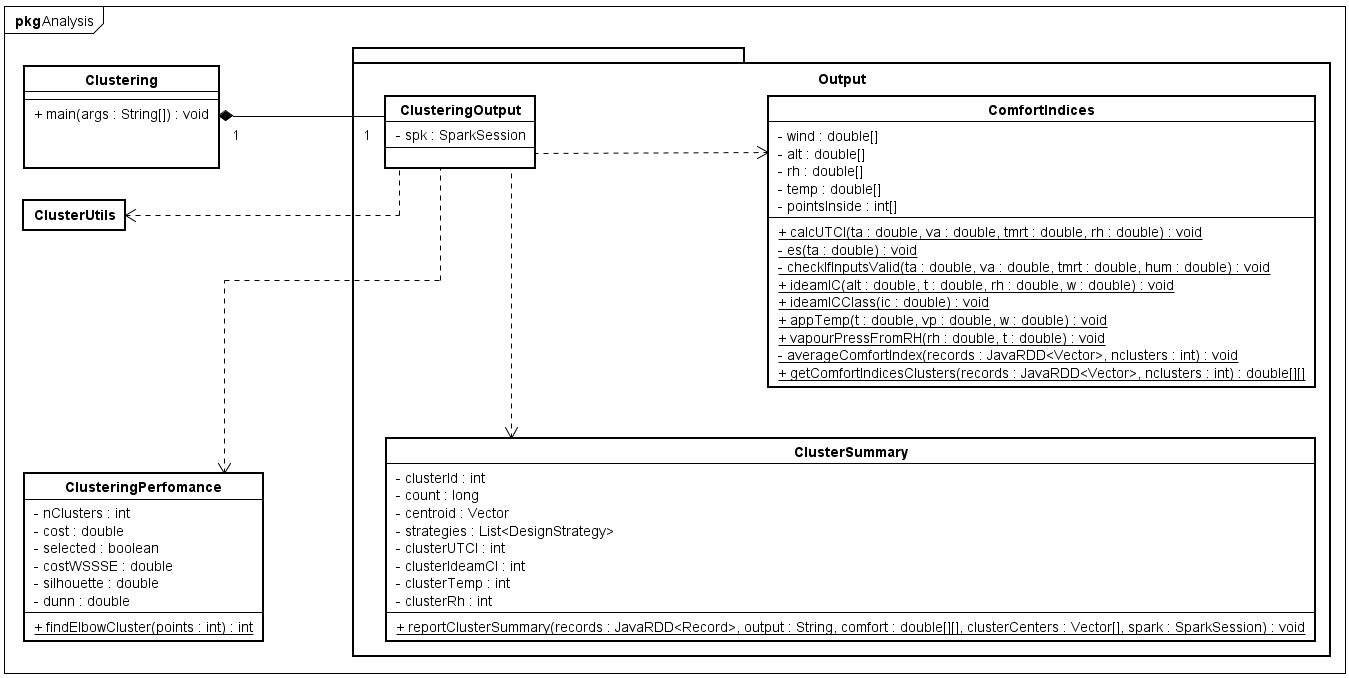


Figure 40 Analysis package and Output sub-package

#### Year + month + cluster

This form of query iterates each year and each month (Figure 41) storing the filtered data and a summary of the frequency of design strategies at each iteration. One data set for each year and each month of the original data is stored. Additionally, for each cluster the frequency of found design strategies is also stored.

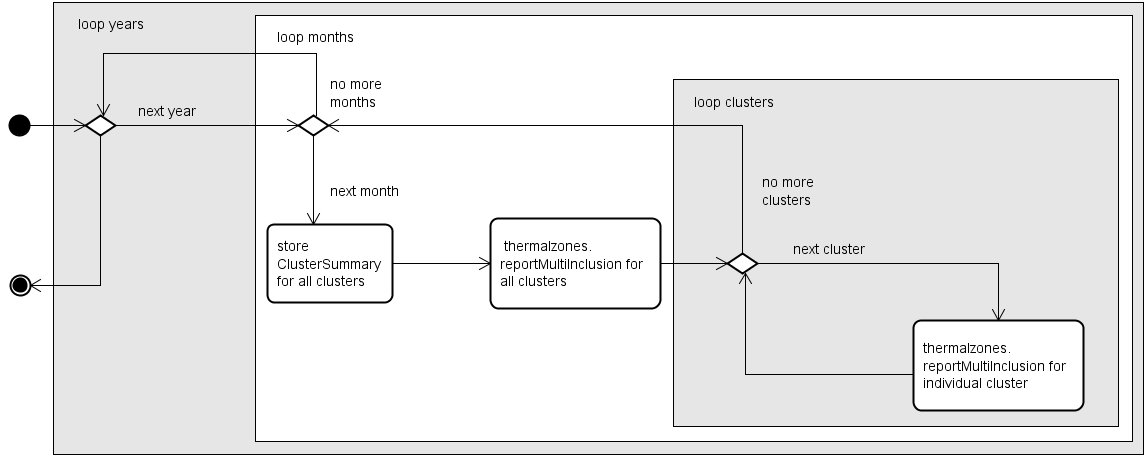


Figure 41 Year + month + cluster

## Dashboard + visualisation

Once the analysis successfully completes the Step on the EMR cluster is given the status of COMPLETED. The UI’s resources monitoring function detects the status and changes the controls available on the interface, enabling the results button and disabling other controls. Clicking the dashboard button triggers an event in the GUIWorkflowBuilder class that checks if a dashboard has been generated, if it exists, the web engine in the UI’s visualise tab is updated with the workflow’s dashboard URL. The application also opens the default browser and points to the same URL.

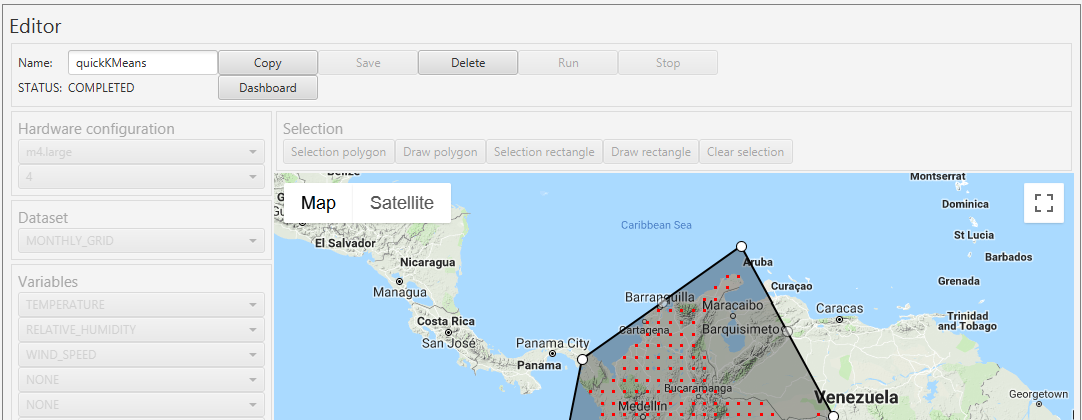


Figure 42 Dashboard button available

If a dashboard does not exist a new folder name is created, this will be located within a public bucket on S3 and contain the dashboard data files and index.html. The GUIWorkflowBuilder creates a new thread that instantiates the GeoVisualisation class (Figure 43). The role of the GeoVisualisation class is to combine files output by the Spark application and transfer the merged files to a newly created space on a public bucket on S3. The Spark application is distributed across several machine instances on EMR specified by the user, any process that requires writing output is also distributed so multiple part files are generated. The GeoVisualisation class iterates all objects with a keypath prefix corresponding to the results location on S3, that was passed to the analysis application. All files that share the same root keypath and include the string “part” in their name are combined and transferred to the public bucket with a new keypath comprised of the dashboard folder name and the file’s root keypath. The text file containing the serialised description of the original workflow is renamed parameters.txt and transferred together with an index.html file. Once the file combine-and-transfer is complete the user can click the dashboard button and examine the dashboard either within the visualise tab in the application on the default browser.

The index.html is a landing page that references a set of shared JavaScript files that define the dashboard framework. Some of these are JavaScript classes while others (shown in grey in Figure 43) contain functions defining the behaviour of the interface. The classes define charts and maps all use the d3.js library (Bostock, 2017a), all the classes are instantiated several times on the dashboard to visualise the results according to the structure defined by the analytic process. Figure 44 describes the sequences of functions called on opening the page and where object instances are created. The zone shaded grey indicates a subset of the behaviour that is triggered when the user interacts with a control on the dashboard. Four sections are provided for examining the results; overview, cluster summary, cluster explorer and comfort comparison.

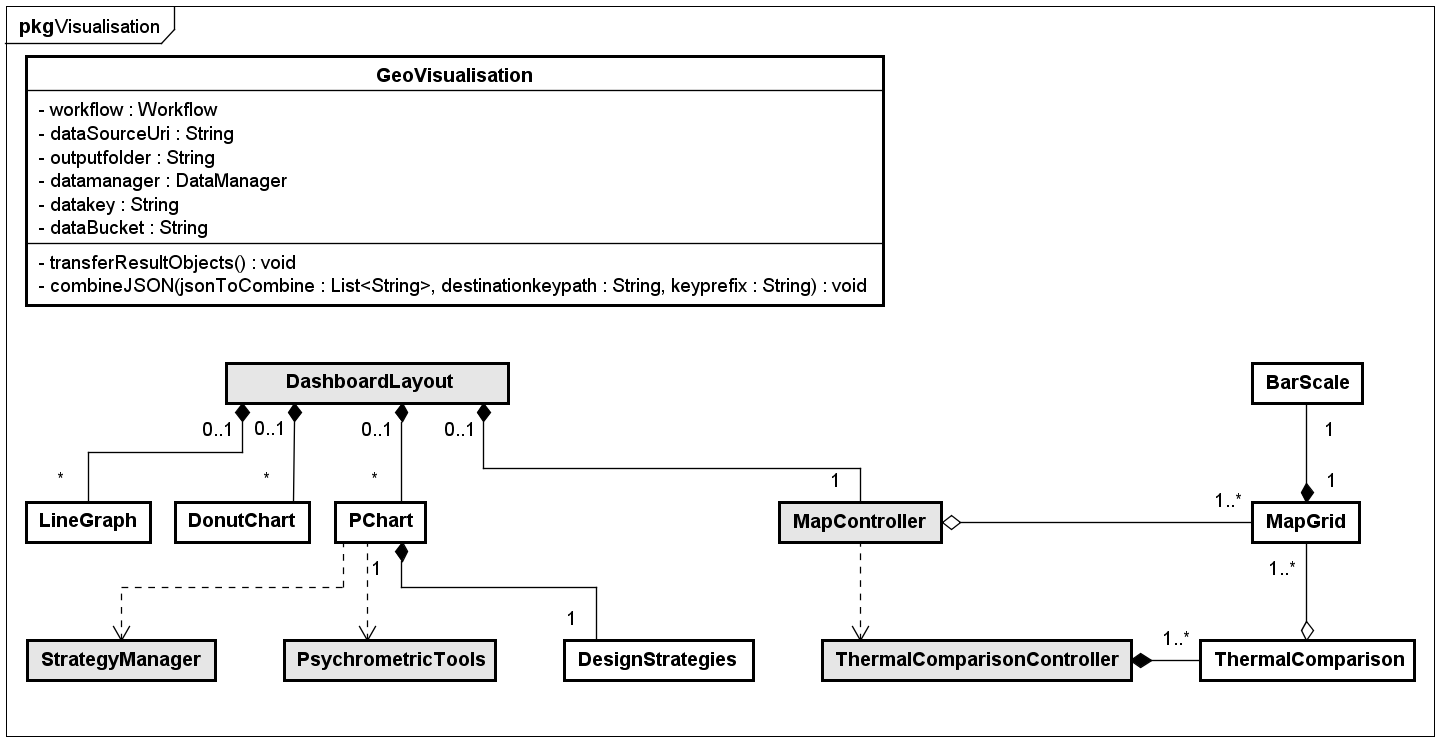


Figure 43 Visualisation package and classes.

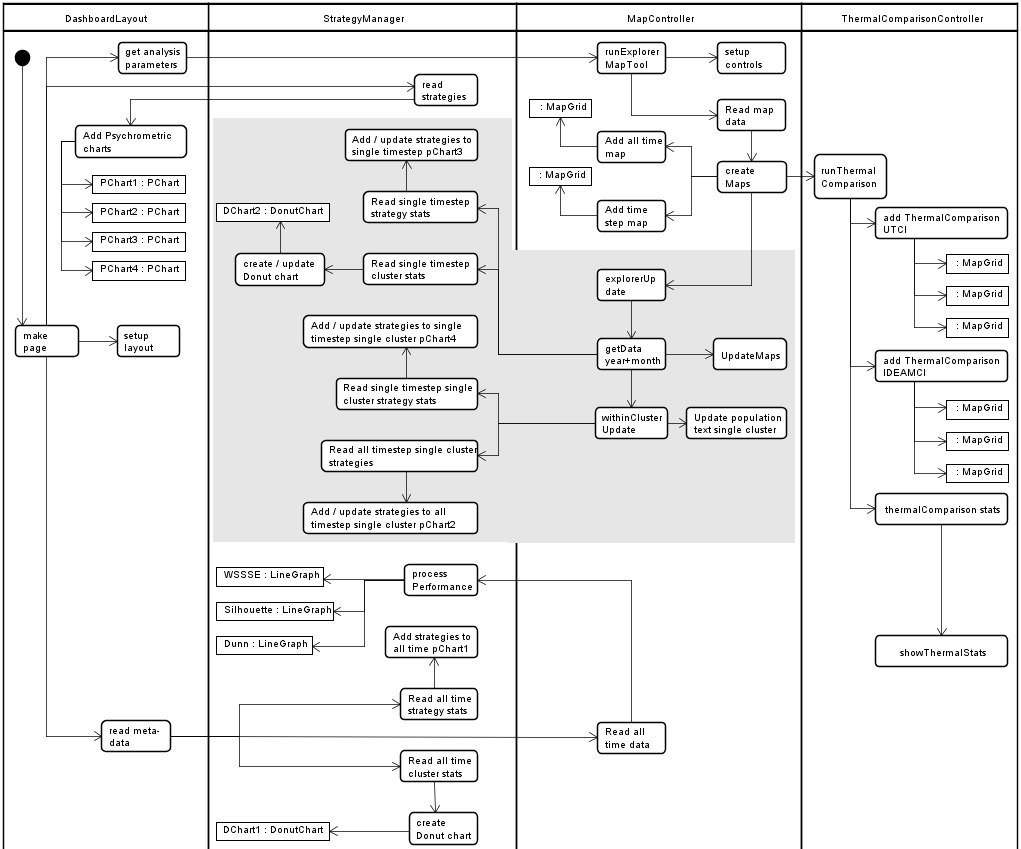


Figure 44 Dashboard behaviour model.

### Overview

The overview section displays the input parameters and performance indices plus the URL of the page. If the analysis involved the optimisation routine LineGraph objects are created for each of the three indices are shown, each with a red dot showing the optimal number of clusters (Figure 45). Scores for each index are shown on the left-hand panel. If the user specified the number of clusters, the score for each index is shown in each panel (Figure 46).



Figure 45 Dashboard overview section for optimisation

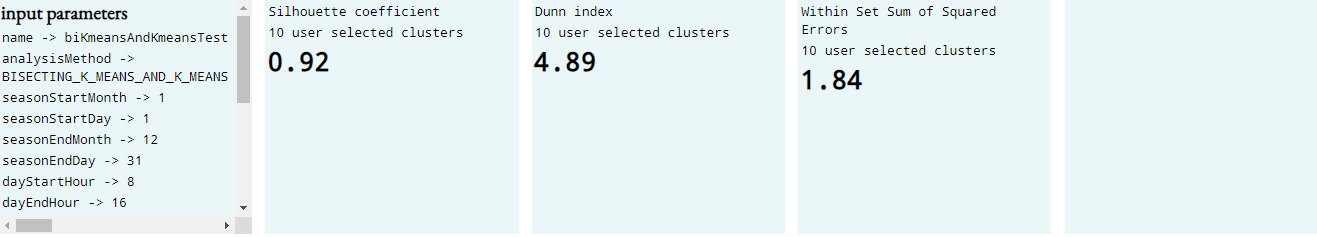


Figure 46 Dashboard overview section for single clustering solution

### Cluster Summary

The summary section is a table summarising the attributes of each cluster defined by the attributes of the ClusterSummary class described in section 7.4.4.1. This includes ranges of the key comfort variables, comfort indices, cluster id, population and associated strategies (Figure 47).

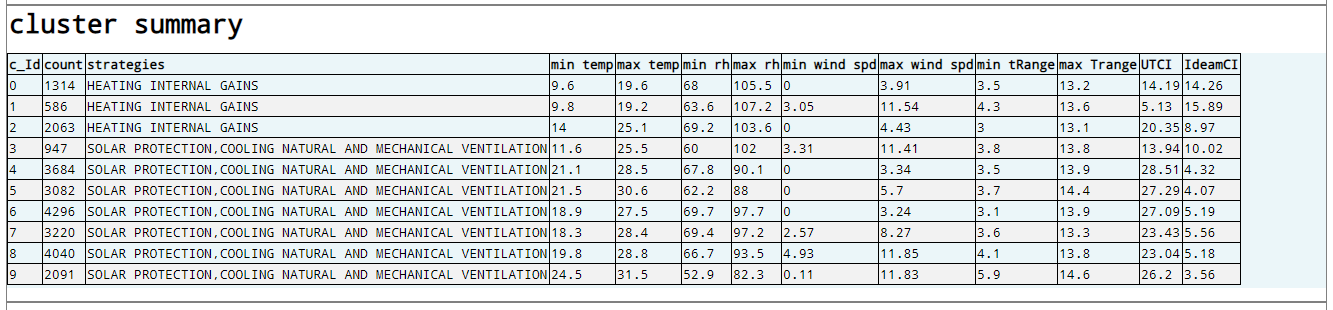


Figure 47 Dashboard summary section

### Cluster explorer

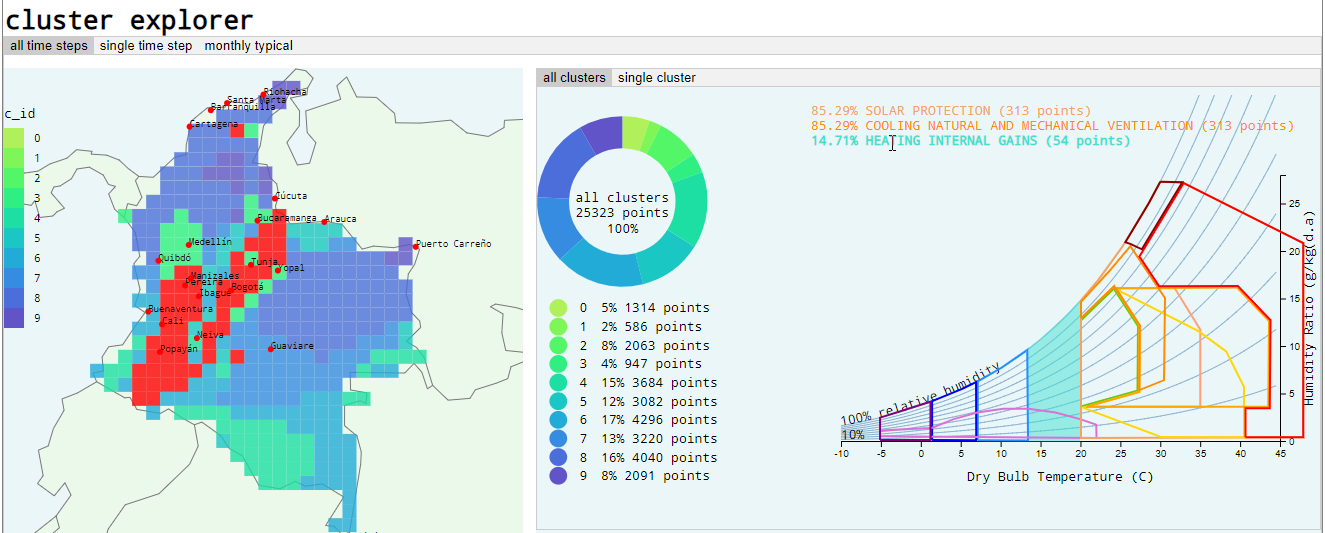


Figure 48 Dashboard cluster explorer all time steps / all clusters with interaction

The cluster explorer section provides various ways of looking at the results. These are accessed by clicking on a tab to select the temporal method and then on the nested tab to select one of the following time-cluster combinations:

1. all time steps / all clusters
2. all time steps / single cluster
3. single time step / all clusters
4. single time step / single cluster
5. monthly typical / all clusters
6. monthly typical / single cluster

On opening the page all time steps / all clusters are presented (Figure 47) this corresponds with the output described in section 7.4.4.2. Instances of the MapGrid, DonutChart and PChart objects are used to provide the user can interact with these to explore the data. The PChart includes DesignStrategy class boundary zones as shown in the UI (Figure 33), listed on the chart are the strategies that correspond with the clusters shown on the MapGrid. Placing the mouse over the text highlights locations on the map where they are applicable and highlights the corresponding boundary on the psychrometric chart (Figure 47). The distribution of points between clusters is shown by the donut chart, placing the mouse over this highlights the corresponding cluster on the map and scale. Similarly, mouse over the map or scale highlights the corresponding clusters on the donut and scale or map. On selecting the single cluster tab, an individual cluster can be selected with a range input, its population, associated strategies and their spatial distribution can be examined (Figure 48).



Figure 49 Dashboard cluster explorer all time steps / single cluster with interaction

Single time step / all clusters mode provides the option of exploring a specific year and month and examining the design strategies associated with that period. The interface includes MapGrid, DonutChart and PChart objects plus range inputs for year and month. This corresponds with the output described in section 7.4.4.4. Selecting the single cluster tab in this mode provides more detailed information of a single cluster at a specific time (Figure 50). Monthly typical / all clusters tab displays the results extracted using the process described in section 7.4.4.3.

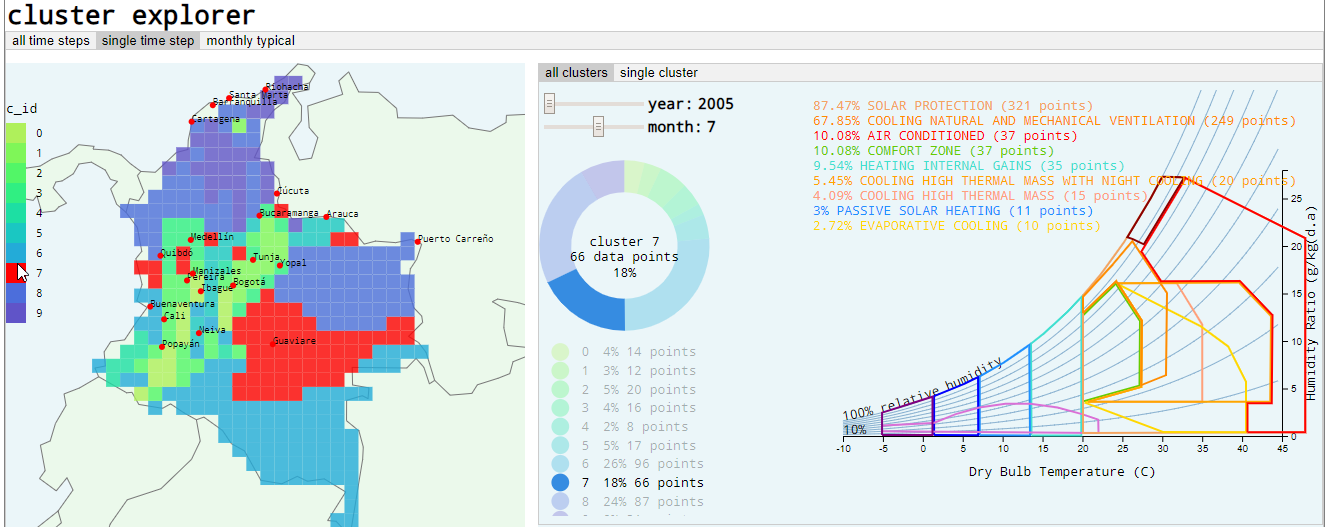


Figure 50 Dashboard cluster explorer single time step / all clusters with interaction

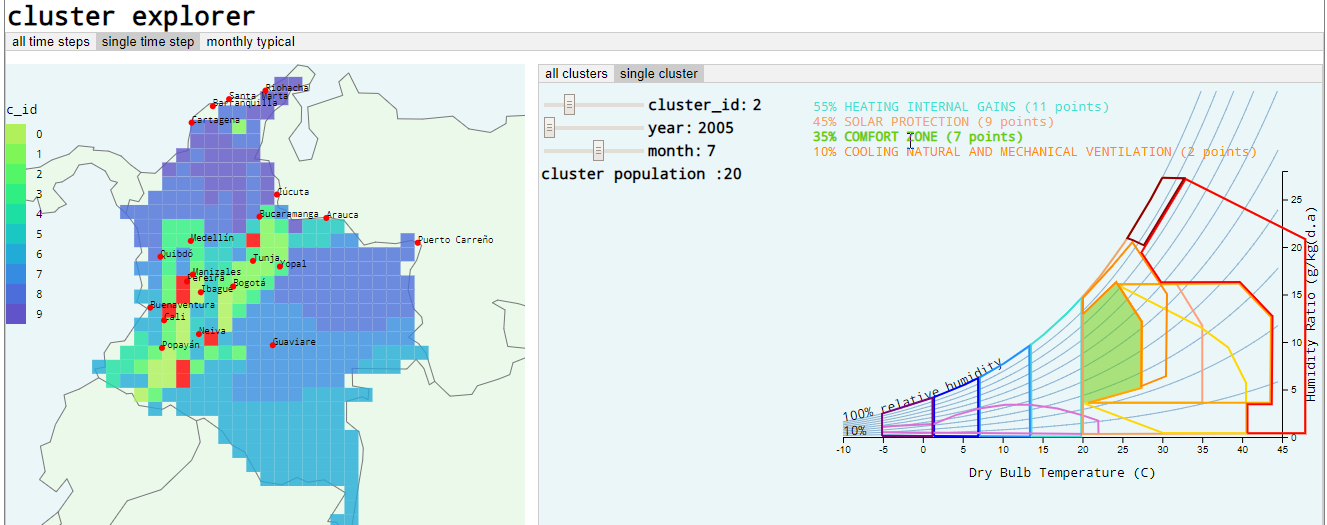


Figure 51 Dashboard cluster explorer single time step / single cluster with interaction

### Comfort comparison

The last section on the dashboard compares comfort indices predicted by the clusters to those calculated by the variables for each grid cell or city (Figure 49). Two comfort indices were defined during the analysis; UTCI and IDEAMCI. For each index a brief description of each index and link to more detail information is provide. A ThermalComparison object is created for each index, this includes three MapGrid objects showing the predicted index, calculated and differences. On the right-hand panel summary statistics of differences between predicted and calculated values is show as mean, standard deviation and root mean squared error.

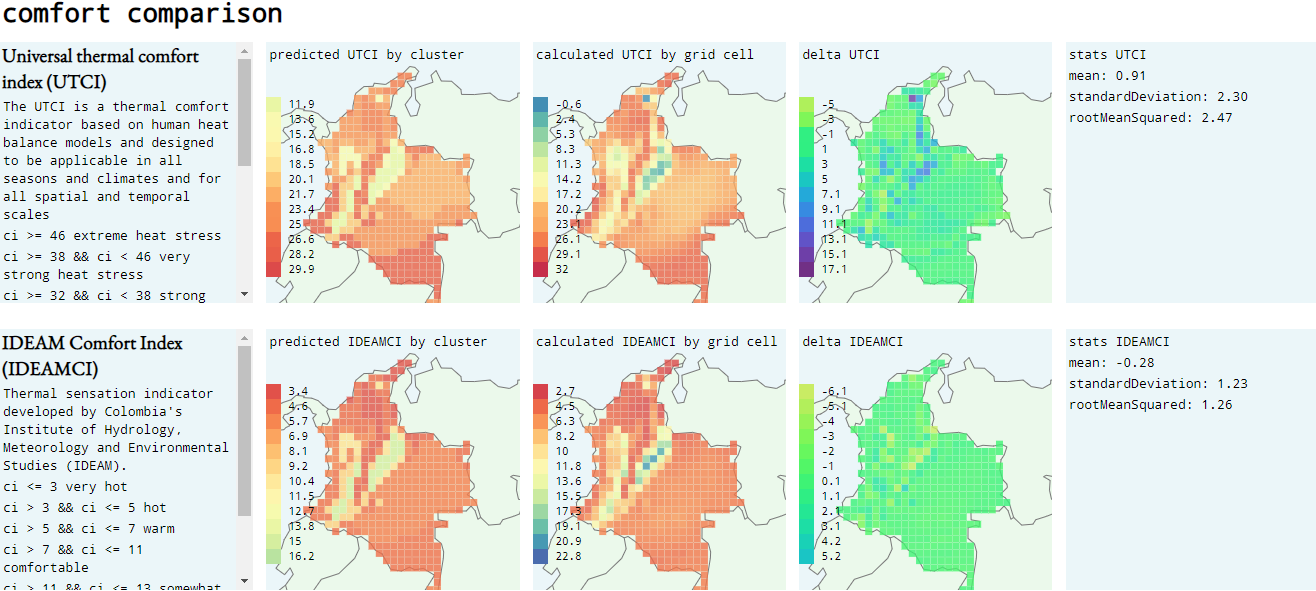


Figure 52 Dashboard thermal index comparison

# Results and evaluation

## Evaluation

The system is intended to produce recommendations for environmental design strategies based on the analysis of a set of weather data. The hypothesis to be tested is that design strategies can be linked with patterns discovered in weather data at various spatiotemporal scales and with different subsets of variables. It is proposed that this can be tested using a big data architecture that enables data analytics over large sets of weather data. Colombia was chosen for a test case as the tropical climate combined with the extreme topography and proximity to oceans. The system is decomposed for evaluation; weather patterns will be assessed both quantitively and qualitatively and matching of recommended design strategies and the patterns evaluated qualitatively.

## Quantitative evaluation of the individual knowledge discovery methods

### Metrics used for evaluation

The literature review identified three validaity metrics to express the dissimilarity (Silhouette) combined cohesion and separation (Dunn) and internal cluster cohesion (WSSSE). The first two were implemented within the analytic process and the latter is already part of Spark’s MLlib. Domain specific quantitative evaluation was implemented using two comfort indices. The difference between the comfort index predicted by the cluster and the index as determined by each data point indicates how well the clusters represent conditions of thermal comfort.

### Evaluation approach

Modelled on a framework for a recent clustering study (Netzel and Stepinski, 2016) a comparison between a series of clustering experiments analysed by the system was undertaken. Based on discussion with domain specialists (section 5.3) and results from Netzel and Stepinski's (2016) study four parameters sets used to define the features for clustering. Combinations of temperature, daily temperature range, relative humidity and wind speed are defined to examine the influence of the wind and the daily temperature range.

1. Temperature, relative humidity (TRh).
2. Temperature, daily temperature range, relative humidity and wind speed (TrRh).
3. Temperature, relative humidity and wind speed (TRhWs).
4. Temperature, daily temperature range, relative humidity and wind speed (TTrRhWs).

The number of clusters to study, k, was determined by running three analytics jobs that sought to optimise k for each of the three methods using the WSSSE validity metric. The hybrid analytic method version, combining hierarchical and non-hierarchical clustering produced identical results to the hierarchal method and was discarded. For the remaining two methods, Silhouette and WSSSE indices of the remaining methods indicated that *k=11* performed well. The Dunn index of the non-hierarchal indicated that *k=6* was optimal. Bisecting k-means (BKM) is the hierarchical method and k-means (KM) the non-hierarchical which gives 16 different analytic workflows (Table 2)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| method | KM | KM | KM | KM | KM | KM | KM | KM |
| k | 6 | 6 | 6 | 6 | 11 | 11 | 11 | 11 |
| features | TRh | TrRh | TRhWs | TTrRhWs | TRh | TrRh | TRhWs | TTrRhWs |
|  |  |  |  |  |  |  |  |  |
|  | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| method | BKM | BKM | BKM | BKM | BKM | BKM | BKM | BKM |
| k | 6 | 6 | 6 | 6 | 11 | 11 | 11 | 11 |
| features | TRh | TrRh | TRhWs | TTrRhWs | TRh | TrRh | TRhWs | TTrRhWs |

Table 2 Evaluation workflows

Each workflow was run on an AWS EMR cluster set up with a r4.xlarge master instance and 9 slave nodes, this configuration was found to perform well when running the initial optimisation studies and AWS . Results from each study are summarised in terms of the performance metrics, comparison to existing comfort indices, processing time and financial cost. The twelve are also compared to each other using dissimilarity matrices based on each of the three validity scores.

### Statistical evaluation of analyses

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **name** | **total secs** | **$** | **sil** | **dunn** | **wssse** | **utci mean** | **utic rmse** | **ideamci mean** | **ideamci mse** |
| KM6TRh | 1716 | 0.319 | 0.97 | 3.53 | 7.6 | 1.24 | 3.35 | -0.14 | 1.36 |
| KM6TrRh | 1991 | 0.371 | 0.97 | 4 | 1.98 | 1.91 | 5.61 | -0.64 | 3.23 |
| KM6TRhWs | 2684 | 0.500 | 0.87 | 2.92 | 34.45 | 0.94 | 3 | -0.26 | 1.47 |
| KM6TTrRhWs | 2726 | 0.507 | 0.77 | 2.85 | 51.21 | 0.98 | 2.99 | -0.24 | 1.46 |
| BKM6TRh | 1648 | 0.307 | 0.97 | 3.8 | 7.63 | 1.41 | 3.35 | -0.24 | 1.36 |
| BKM6TrRh | 2602 | 0.484 | 0.93 | 2.65 | 2.39 | 1.93 | 5.63 | -0.67 | 3.25 |
| BKM6TRhWs | 2740 | 0.510 | 0.84 | 2.78 | 36.04 | 0.84 | 2.88 | -0.26 | 1.37 |
| BKM6TTrRhWs | 2719 | 0.506 | 0.77 | 2.56 | 54.25 | 1.06 | 3.35 | -0.33 | 1.74 |
| KM11TRh | 2623 | 0.488 | 0.99 | 5.84 | 2.51 | 1.28 | 3.18 | -0.16 | 1.27 |
| KM11TrRh | 2660 | 0.495 | 0.99 | 8.46 | 0.68 | 1.93 | 5.62 | -0.65 | 3.23 |
| KM11TRhWs | 4242 | 0.789 | 0.93 | 4.4 | 19.01 | 0.49 | 2.4 | -0.09 | 1.29 |
| KM11TTrRhWs | 4061 | 0.756 | 0.88 | 4.13 | 34.83 | 0.6 | 2.77 | -0.11 | 1.39 |
| BKM11TRh | 2624 | 0.488 | 0.99 | 4.14 | 2.8 | 1.33 | 3.17 | -0.2 | 1.3 |
| BKM11TrRh | 3918 | 0.729 | 0.98 | 5.04 | 0.83 | 1.93 | 5.61 | -0.66 | 3.24 |
| BKM11TRhWs | 4128 | 0.768 | 0.9 | 2.93 | 21.43 | 0.65 | 2.17 | -0.18 | 1.27 |
| BKM11TTrRhWs | 2500 | 0.465 | 0.82 | 2.92 | 38.07 | 0.71 | 2.85 | -0.23 | 1.6 |

No single method tested showed best performance across all indices

Method

Top 4 performing methods were equally populated by BKM and KM, but KM took the top spot in all indices

K clusters

11 clusters performed better across all indices

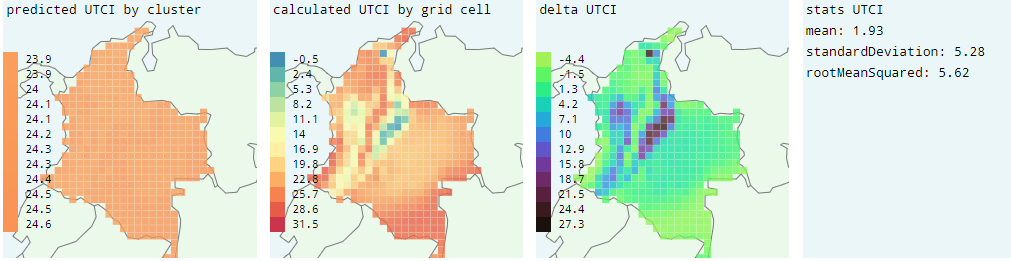
Parameters

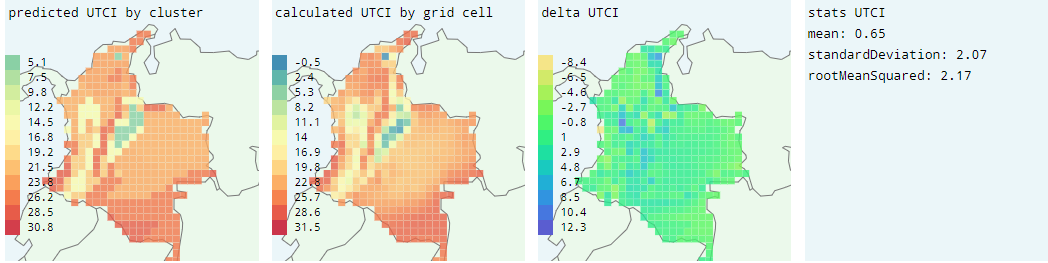
Comfort comparison show that the methods failed to capture the colder extremes – estimating warmer conditions in both (IDEAMCI lower values are hotter conditions)

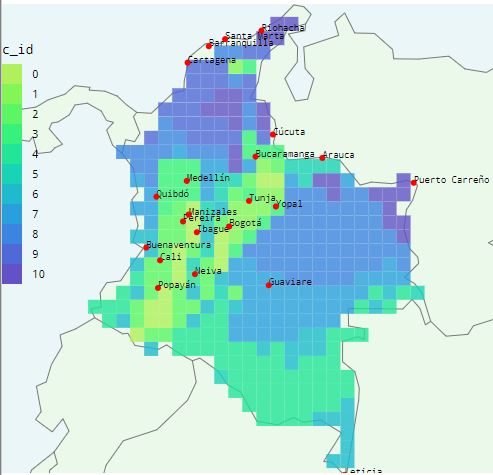
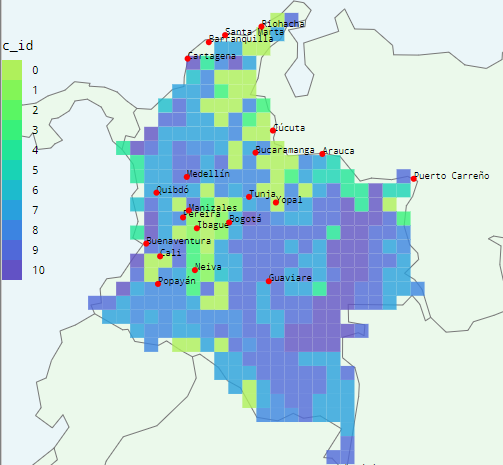
Including temperature range tended to result in poorer performance in terms of the comfort comparison but better scores in WSSSE. Wind speed was important in scoring well in terms of comfort comparison but made little impact on the silhouette index

WSSSE should be minimised but analyses with small WSSSE performed poorly in comfort comparisons

### Graphical evaluation

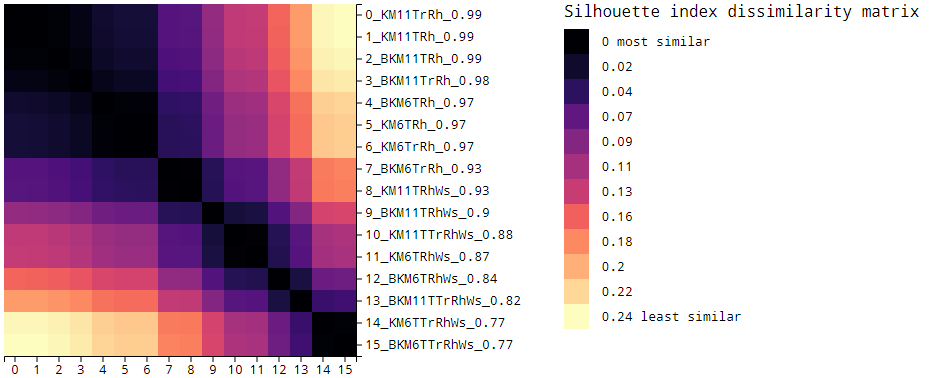


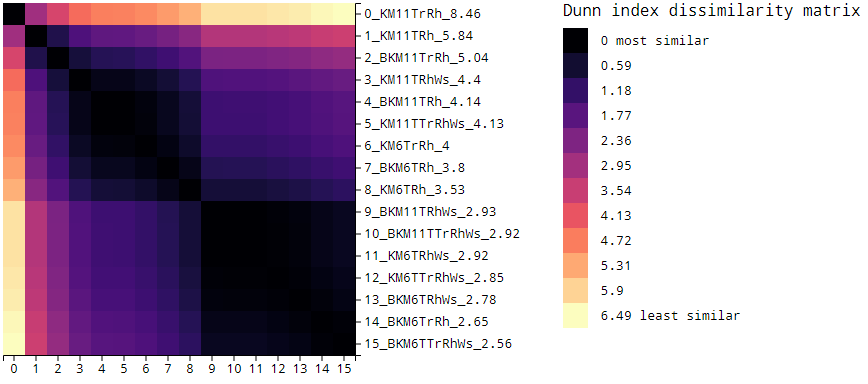




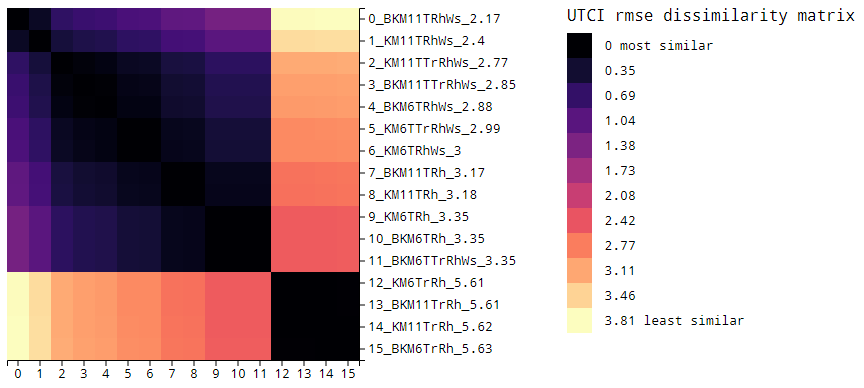
Including Temperature range failed to create clusters that reflected the topography – an indicator of climate variation in Colombia (see next section)

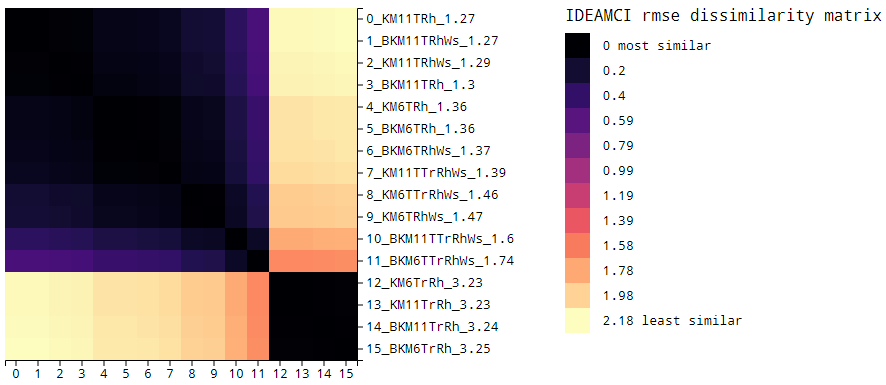
### Comparison between analyses











## Qualitative evaluation

### Opinion by domain experts

A software walkthrough and a brief (Hudson, 2018) project background was presented to a group of trained architects based in Colombia and the UK, which included academic researchers and practicing architects. All participants had either worked with or research concepts of thermal comfort and low-energy design and represented domain experts.

One major difficulty describing aspects of the project relating to the spatiotemporal data and its multi-dimensionality.

The notion of data exploration was hard to covey – they were people looking for

analytic methods and the data literacy of the participants, how clustering worked concepts of data vectors with features and cloud architecture of the project

Only a single set of results was discussed – the notion of data exploration and knowledge discovery was difficult to convey and perhaps missed completely by certain experts. really requires more than this

The authors view of the anticipated user was an overestimation see as a diligent and determined designer keen to ensure designs were consuming minimal energy as comfortable for occupants and providing users a unique experience in response to the built spaces, the buildings context and the local climate.

According Colombian participants to the typical user simple wants to meet all regulatory goals and finish the work as quickly as possible

# Conclusions

## Lessons Learned

New technology Eclipse IDE, Maven, AWS SDK, Spark, JavaRDD, Dataframes, JavaFx GMapsFX

Computing on EMR clusters how to debug them

Data manipulation on S3

AMDD test driven

Unit testing

Integration tests

JSON with Jackson library

Need to switch form the RDD MLlib to the Dataframe version

Difficulties of the spatiotemporal data

Both visualising results – anticipating the slicing of data, and organisaitng the original data

## Knowledge discovery

The application and provides KD Domain expertise, data management and analytics combined with easy ways to interact, explore and analyse data

Data is output as JSON formated anticpating other

But could be improved by

## Evaluation results

Comments from experts

Statistical comparison

## Future research

More extensive study of different analytic methods,

Methods for measuring validity

Deeper understanding of the analytic method implications Literature on Clustering with climate data is vast many studies show use if different metrics for validity and more complex workflows involving other machine learning methods to both classify and validate. See Netzel paper DTW metric for comparing dissimilarity in time series.

Deeper understanding of the variable selection implications Extensive study of the best way to selected and pre-process variables – here the choice was left in the hands of the expert which may not be entirely inaccurate. PCA

Workflow is largely fixed in terms of the units of work which are configured by specific parameters, future research could look at how to combine analytic methods

Significance testing on results

Domain usage – how it will be used in practice

Time series

ST data in analysis and representation

# References

Ambler, S. W. (2004) *The object primer : agile modeling-driven development with UML 2.0*. Cambridge University Press.

Arbelaitz, O. *et al.* (2013) ‘An extensive comparative study of cluster validity indices’, *Pattern Recognition*. Pergamon, 46(1), pp. 243–256. doi: 10.1016/J.PATCOG.2012.07.021.

ASHRAE (2013) *2013 ASHRAE Handbook: Fundamentals*, *ASHRAE*. doi: 10.1163/ej.9789004155947.i-937.23.

Atluri, G., Karpatne, A. and Kumar, V. (2017) ‘Spatio-Temporal Data Mining: A Survey of Problems and Methods’, *ACM Comput. Surv*, 1(1). doi: 10.1145/nnnnnnn.nnnnnnn.

Avci Salma, C., Tekinerdogan, B. and Athanasiadis, I. N. (2017) ‘Chapter 4 – Domain-Driven Design of Big Data Systems Based on a Reference Architecture’, in *Software Architecture for Big Data and the Cloud*, pp. 49–68. doi: 10.1016/B978-0-12-805467-3.00004-1.

*AWS Toolkit for Eclipse* (no date). Available at: https://aws.amazon.com/eclipse/ (Accessed: 3 October 2018).

Begoli, E. and Horey, J. (2012) ‘Design principles for effective knowledge discovery from big data’, in *Proceedings of the 2012 Joint Working Conference on Software Architecture and 6th European Conference on Software Architecture, WICSA/ECSA 2012*, pp. 215–218. doi: 10.1109/WICSA-ECSA.212.32.

Bostock, M. (2017a) *D3.js - Data-Driven Documents*. Available at: https://d3js.org/ (Accessed: 28 June 2018).

Bostock, M. (2017b) *TopoJSON*, *Github*. Available at: https://github.com/topojson/topojson (Accessed: 3 May 2017).

Bracco, A. *et al.* (2017) ‘Advancing climate science with knowledge-discovery through data mining’, *npj Climate and Atmospheric Science*, 1(1), p. 4. doi: 10.1038/s41612-017-0006-4.

Buyya, R. *et al.* (2016) ‘Chapter 18 – eScience and Big Data Workflows in Clouds: A Taxonomy and Survey’, in *Big Data*, pp. 431–455. doi: 10.1016/B978-0-12-805394-2.00018-0.

Cgiar-csi.org. (2012) *CRU-TS v3.10.01 Historic Climate Database for GIS | CGIAR-CSI.* Available at: http://www.cgiar-csi.org/data/uea-cru-ts-v3-10-01-historic-climate-database (Accessed: 5 November 2017).

Chouksey, P. and Chauhan, A. S. (2017) ‘Weather Data Analytics using MapReduce and Spark’, *International Journal of Advanced Research in Computer and Communication Engineering*, 6(2). doi: 10.17148/IJARCCE.2017.6210.

Dagade, V. *et al.* (2015) ‘Big Data Weather Analytics Using Hadoop’, *International Journal of Emerging Technology in Computer Science & Electronics*, 14(2), pp. 976–1353. Available at: https://pdfs.semanticscholar.org/f2e4/918444be9b30f29132e93ce02d29ccf26eda.pdf (Accessed: 29 May 2018).

Das, T., Zaharia, M. and Wendell, P. (2015) *Spark Streaming*. Available at: https://databricks.com/blog/2015/07/30/diving-into-apache-spark-streamings-execution-model.html (Accessed: 21 February 2018).

Degaetano, A. T. (1996) ‘Delineation of Mesoscale Climate Zones in the Northeastern United States Using a Novel Approach to Cluster Analysis’, *Journal of Climate*, 9(8), pp. 1765–1782. doi: 10.1175/1520-0442(1996)009<1765:DOMCZI>2.0.CO;2.

Dimoudi, A. and Tompa, C. (2008) ‘Energy and environmental indicators related to construction of office buildings’, *Resources, Conservation and Recycling*, 53(1–2), pp. 86–95. doi: 10.1016/j.resconrec.2008.09.008.

Faghmous, J. H. and Kumar, V. (2014) ‘Spatio-temporal Data Mining for Climate Data: Advances, Challenges, and Opportunities’, in Chu, W. (ed.) *Springer Berlin Heidelberg*, pp. 83–116. doi: 10.1007/978-3-642-40837-3\_3.

Fang, W. *et al.* (2014) ‘Meteorological data analysis using MapReduce.’, *The Scientific World Journal*, 2014, p. 646497. doi: 10.1155/2014/646497.

El Fazziki, A. *et al.* (2015) ‘A multi-agent framework for a hadoop based air quality decision support system’, in *CEUR Workshop Proceedings*, pp. 45–59.

Forsythe, N., Blenkinsop, S. and Fowler, H. J. (2015) ‘Exploring objective climate classification for the Himalayan arc and adjacent regions using gridded data sources’, *Earth System Dynamics*, 6(1), pp. 311–326. doi: 10.5194/esd-6-311-2015.

Fountalis, I., Bracco, A. and Dovrolis, C. (2014) ‘Spatio-temporal network analysis for studying climate patterns’, *Climate Dynamics*, 42(3–4), pp. 879–899. doi: 10.1007/s00382-013-1729-5.

Fovell, R. G. and Fovell, M. Y. C. (1993) ‘Climate zones of the conterminous United States defined using cluster analysis’, *Journal of Climate*, pp. 2103–2135. doi: 10.1175/1520-0442(1993)006<2103:CZOTCU>2.0.CO;2.

Givoni, B. (1992) ‘Comfort, climate analysis and building design guidelines’, *Energy and Buildings*, 18, pp. 11–23. Available at: https://ac-els-cdn-com.liverpool.idm.oclc.org/037877889290047K/1-s2.0-037877889290047K-main.pdf?\_tid=67644909-7d5e-4718-b47f-80732d024251&acdnat=1527603110\_130b2a2e77cd97d48361202940532374 (Accessed: 29 May 2018).

González, O. (1998) *Metodología para el Calculo del Confort Climático en Colombia: NOTA TECNICA DEL IDEAM.* Bogota. Available at: http://documentacion.ideam.gov.co/openbiblio/bvirtual/007574/Metodologiaconfort.

Höppe, P. (1999) ‘The physiological equivalent temperature - A universal index for the biometeorological assessment of the thermal environment’, *International Journal of Biometeorology*. doi: 10.1007/s004840050118.

*http-server: a command-line http server* (2018) *npmjs*. Available at: https://www.npmjs.com/package/http-server (Accessed: 3 October 2018).

Hudson, R. (2018) *Big Climate Data Analytics: Effective Knowledge Discovery from Colombia’s Weather Data*. Available at: https://docs.google.com/presentation/d/1QAY6ZUAJ5zlwWwdgNoAhxZlYenV0TnsVtp8n6fcSLdM/edit?usp=sharing.

Jain, A. K. (2010) ‘Data clustering: 50 years beyond K-means’, *Pattern Recognition Letters*. North-Holland, 31(8), pp. 651–666. doi: 10.1016/J.PATREC.2009.09.011.

Jain, A. K. and Dubes, R. C. (1988) *Algorithms for clustering data*. Prentice Hall. Available at: https://dl.acm.org/citation.cfm?id=46712 (Accessed: 22 October 2018).

Jain, H. and Jain, R. (2017) ‘Big data in weather forecasting: Applications and challenges’, in *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*. IEEE, pp. 138–142. doi: 10.1109/ICBDACI.2017.8070824.

Jarvis, A., H.I. Reuter, A. Nelson, E. G. (2008) *Hole-filled SRTM 90m for the globe Version 4 Database*. Available at: http://srtm.csi.cgiar.org (Accessed: 5 November 2017).

Jayanthi, D. and Sumathi, G. (2017) ‘Weather data analysis using spark — An in-memory computing framework’, in *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*. IEEE, pp. 1–5. doi: 10.1109/IPACT.2017.8245142.

Jayaratne, M. *et al.* (2017) ‘Apache spark based distributed self-organizing map algorithm for sensor data analysis’, in *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, pp. 8343–8349. doi: 10.1109/IECON.2017.8217465.

Jendritzky, G. and Höppe, P. (2017) ‘The UTCI and the ISB’, *International Journal of Biometeorology*. doi: 10.1007/s00484-017-1390-5.

Jones, P. and Harris, I. (2008) *Climatic Research Unit (CRU) time-series datasets of variations in climate with variations in other phenomena*, *NCAS British Atmospheric Data Centre*. Available at: http://catalogue.ceda.ac.uk/uuid/3f8944800cc48e1cbc29a5ee12d8542d (Accessed: 5 November 2017).

Karpatne, A. *et al.* (2017) ‘Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data’, *IEEE Transactions on Knowledge and Data Engineering*, 29(10), pp. 2318–2331. doi: 10.1109/TKDE.2017.2720168.

*KMeansModel (Spark 2.3.2 JavaDoc)* (no date). Available at: https://spark.apache.org/docs/latest/api/java/org/apache/spark/mllib/clustering/KMeansModel.html (Accessed: 18 October 2018).

Kodinariya, T. M. and Makwana, P. R. (2013) ‘Review on determining number of Cluster in K-Means Clustering’, *International Journal of Advance Research in Computer Science and Management Studies*.

Liu, F. *et al.* (2011) *NIST Cloud Computing Reference Architecture Recommendations of the National Institute of Standards and Technology*. Available at: https://ws680.nist.gov/publication/get\_pdf.cfm?pub\_id=909505 (Accessed: 3 October 2018).

Liu, Y. and Weisberg, R. H. (2005) ‘Patterns of ocean current variability on the West Florida Shelf using the self-organizing map’, *Journal of Geophysical Research: Oceans*, 110(6), pp. 1–12. doi: 10.1029/2004JC002786.

Liu, Y. and Weisberg, R. H. (2011) ‘A Review of Self-Organizing Map Applications in Meteorology and Oceanography’, in Igadwa Mwasiagi, J. (ed.) *Self Organizing Maps - Applications and Novel Algorithm Design*. www.intechopen.com. doi: 10.5772/13146.

Liu, Y., Weisberg, R. H. and Mooers, C. N. K. (2006) ‘Performance evaluation of the self-organizing map for feature extraction’, *Journal of Geophysical Research: Oceans*, 111(5). doi: 10.1029/2005JC003117.

Luna-Romera, J. M. *et al.* (2016) ‘An approach to silhouette and dunn clustering indices applied to big data in spark’, in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/978-3-319-44636-3\_15.

Manzano-Agugliaro, F. *et al.* (2015) ‘Review of bioclimatic architecture strategies for achieving thermal comfort’, *Renewable and Sustainable Energy Reviews*. Pergamon, 49, pp. 736–755. doi: 10.1016/J.RSER.2015.04.095.

Mariam Varghese, S. (2015) ‘Leveraging Map Reduce With Hadoop for Weather Data Analytics’, *IOSR Journal of Computer Engineering Ver. II*, 17(3), pp. 2278–661. doi: 10.9790/0661-17320612.

Mell, P. and Grance, T. (2011) *The NIST Definition of Cloud Computing Recommendations of the National Institute of Standards and Technology*. doi: 10.6028/NIST.SP.800-145.

Milne, M., Liggett, R. and Benson, A. (2009) ‘Climate Consultant 4.0 develops design guidelines for each unique climate’, *American Solar Energy Society Meeting*. Available at: http://www.energy-design-tools.aud.ucla.edu/papers/ases09-milne.pdf (Accessed: 10 April 2018).

*MLlib: Main Guide - Spark 2.3.2 Documentation* (no date). Available at: https://spark.apache.org/docs/latest/ml-guide.html (Accessed: 16 October 2018).

Netzel, P. and Stepinski, T. (2016) ‘On Using a Clustering Approach for Global Climate Classification’, *Journal of Climate*, 29(9), pp. 3387–3401. doi: 10.1175/JCLI-D-15-0640.1.

Nikolaou, T. G. *et al.* (2012) ‘On the application of clustering techniques for office buildings’ energy and thermal comfort classification’, *IEEE Transactions on Smart Grid*. doi: 10.1109/TSG.2012.2215059.

Olgyay, V. and Olgyay, A. (1963) *Design With Climate: Bioclimatic Approach to Architectural Regionalism*. Princeton University Press.

Omer, A. M. (2008) ‘Energy, environment and sustainable development’, *Renewable and Sustainable Energy Reviews*, 12, pp. 2265–2300. doi: 10.1016/j.rser.2007.05.001.

Piech, C. (2013) *K Means Stanford CS221*. Available at: http://stanford.edu/~cpiech/cs221/handouts/kmeans.html (Accessed: 22 October 2018).

Rhee, J. *et al.* (2008) ‘Delineation of climate regions using in-situ and remotely-sensed data for the Carolinas’, *Remote Sensing of Environment*, 112(6), pp. 3099–3111. doi: 10.1016/j.rse.2008.03.001.

Rodenburg, B. and Maria Fiore, M. (2017) *Detecting Weather Twins using Apache Spark*, *LSDE: Large Scale Data Engineering 2017*. Available at: https://event.cwi.nl/lsde/2017/showcase\_n2.shtml (Accessed: 29 May 2018).

Rodriguez, M. A. and Buyya, R. (2017) ‘Chapter 18 – Scientific Workflow Management System for Clouds’, in *Software Architecture for Big Data and the Cloud*, pp. 367–387. doi: 10.1016/B978-0-12-805467-3.00018-1.

Terpilowski, R. (no date) *GMapsFX*. Available at: https://rterp.github.io/GMapsFX/ (Accessed: 18 October 2018).

Thinsungnoen, T. *et al.* (2015) ‘The Clustering Validity with Silhouette and Sum of Squared Errors’, in *The Proceedings of the 2nd International Conference on Industrial Application Engineering 2015*. doi: 10.12792/iciae2015.012.

Unidata (2012) ‘NetCDF-Java library and TDS version 4.6.9’. Boulder. CO: UCAR/Unidata. doi: http://doi.org/10.5065/D6RN35XM.

Wentz, F. *et al.* (2015) *Remote Sensing Systems Cross-Calibrated Multi-Platform (CCMP) 6-hourly ocean vector wind analysis product on 0.25 deg grid, Version 2.0, [subset: CCMP V2.0 Level-3.5].* Santa Rosa, CA: Remote Sensing Systems. Available at: www.remss.com/measurements/ccmp (Accessed: 5 November 2017).

Wiwie, C., Baumbach, J. and Röttger, R. (2015) ‘Comparing the performance of biomedical clustering methods’, *Nature Methods*. doi: 10.1038/nmeth.3583.

Zaharia, M. *et al.* (2013) ‘Discretized Streams: Fault-Tolerant Streaming Computation at Scale’, *Sosp*, (1), pp. 423–438. doi: 10.1145/2517349.2522737.

Zscheischler, J., Mahecha, M. D. and Harmeling, S. (2012) ‘Climate classifications: The value of unsupervised clustering’, in *Procedia Computer Science*, pp. 897–906. doi: 10.1016/j.procs.2012.04.096.

# List of Tables

[Table 1 Key Research methods used in the dissertation 7](#_Toc526246822)

[Table 2 System development process 8](#_Toc526246823)

# List of Figures

[Figure 1 Agile model driven design phases 7](#_Toc527887142)

[Figure 2 Manzano-Agugliaro *et al.* (2015) adapted version of the psychrometric chart 10](#_Toc527887143)

[Figure 3 Reference architecture of a WMS 15](#_Toc527887144)

[Figure 4 Big data reference architecture (Avci Salma, Tekinerdogan and Athanasiadis, 2017) 16](#_Toc527887145)

[Figure 5 Define workflow 20](#_Toc527887146)

[Figure 6 Run and monitor workflow 20](#_Toc527887147)

[Figure 7 Output and visualise results 21](#_Toc527887148)

[Figure 8 Manage design strategies 21](#_Toc527887149)

[Figure 9 Proposed architecture (extended from: Lopes, Palmer and O’Sullivan, 2017) 22](#_Toc527887150)

[Figure 10 Architecture for the system. 22](#_Toc527887151)

[Figure 11 Simplified package diagram. 22](#_Toc527887152)

[Figure 12 Sequence diagram for defining and running a workflow 23](#_Toc527887153)

[Figure 13 Sequence diagram for monitoring workflows and resources. 24](#_Toc527887154)

[Figure 14 Sequence diagram for show or generate dashboard. 24](#_Toc527887155)

[Figure 15 Sequence diagram for generate dashboard also see Figure 14. 25](#_Toc527887156)

[Figure 16 Development increments (extended from Ambler (2004), p119). 25](#_Toc527887157)

[Figure 17 Deployment on AWS 27](#_Toc527887158)

[Figure 18 Flow diagram showing the stages of the data preparation process. 28](#_Toc527887159)

[Figure 19 Coordination, User Interface and Visualisation packages overview 29](#_Toc527887160)

[Figure 20 Analysis package overview 30](#_Toc527887161)

[Figure 21 Workflowbuilder package class diagrams a 31](#_Toc527887162)

[Figure 22 Workflowbuilder package class diagrams b 32](#_Toc527887163)

[Figure 23 Coordination package class diagrams 33](#_Toc527887164)

[Figure 24 GUI for Workflowbuilder 34](#_Toc527887165)

[Figure 25 Resource and workflow monitoring algorithm 35](#_Toc527887166)

[Figure 26 UI Monitor resources and workflows tab 36](#_Toc527887167)

[Figure 27 Workflow state machine 36](#_Toc527887168)

[Figure 28 Workflow with status COMPLETED 37](#_Toc527887169)

[Figure 29 Processing workflows activity diagram 37](#_Toc527887170)

[Figure 30 Run workflow activity diagram 37](#_Toc527887171)

[Figure 31 Design strategy editor mock-up. 38](#_Toc527887172)

[Figure 32 Parameters for a Step on EMR 39](#_Toc527887173)

[Figure 33 Analysis package and ReadFilter sub-package. 39](#_Toc527887174)

[Figure 34 Analytics overview activity diagram 40](#_Toc527887175)

[Figure 35 Analysis package and Performance sub-package 41](#_Toc527887176)

[Figure 36 Elbow method 42](#_Toc527887177)

[Figure 37 Analysis package and Classification sub-package 42](#_Toc527887178)

[Figure 38 Analysis package and Output sub-package 44](#_Toc527887179)

[Figure 39 Year + month + cluster 45](#_Toc527887180)

[Figure 40 Dashboard button available 45](#_Toc527887181)

[Figure 41 Visualisation package and classes. 46](#_Toc527887182)

[Figure 42 Dashboard behaviour model. 47](#_Toc527887183)

[Figure 43 Dashboard overview section for optimisation 47](#_Toc527887184)

[Figure 44 Dashboard overview section for single clustering solution 48](#_Toc527887185)

[Figure 45 Dashboard cluster explorer all time steps / all clusters with interaction 48](#_Toc527887186)

[Figure 46 Dashboard cluster explorer all time steps / single cluster with interaction 49](#_Toc527887187)

[Figure 47 Dashboard cluster explorer single time step / all clusters with interaction 49](#_Toc527887188)

[Figure 48 Dashboard cluster explorer single time step / single cluster with interaction 50](#_Toc527887189)

[Figure 49 Dashboard thermal index comparison 50](#_Toc527887190)

[Figure 50 User Interface package 51](#_Toc527887191)

1. Source code is available: <https://github.com/rolyhudson/climacolombia.git>. [↑](#footnote-ref-1)