

|  |
| --- |
| Coventry University  5011CEM Big Data Programming Project Specification Document  Ensemble comparisons and simple visualization |

**Author: Armandas Barkauskas**

**Document Version: 1.1**

**10/02/2020**

**Table of Contents**

[1. Introduction 3](#_Toc30950202)

[2. Project Requirements 3](#_Toc30950203)

[2.1 Related documents 3](#_Toc30950204)

[2.2 Terms/Acronyms and Definitions 3](#_Toc30950205)

[3. Risks and Assumptions 3](#_Toc30950206)

[4. Out of Scope 3](#_Toc30950207)

[5. System/ Solution Overview 3](#_Toc30950208)

[5.1 Context Diagram/ Interface Diagram/ Data Flow Diagram, Application Screen Flow, Sitemap, Process Flow 3](#_Toc30950209)

[6. Project Management 4](#_Toc30950210)

[7. References 4](#_Toc30950211)

[8. Open Issues 4](#_Toc30950212)

[Appendix 4](#_Toc30950213)

# Introduction

This aim of this project is to use multiple observation models for Europe area and provide more accurate results. The first approach of doing this is simple mean ensemble, which is basically taking an average of all models combined for each location point. According to, Hyde, it was found that the accuracy of model ensembles can be improved up to 18% when using the DDC clustering algorithm (Hyde and Angelov 2014). Project will involve using the sub-spacing and parallel processing concepts as running big data processes sequentially takes a long time. Ultimately, both ensembles and comparison results will be plotted in a simple map for a visualisation part. More information about DDC algorithm can be found at <https://ieeexplore.ieee.org/document/6930157>.

# Project Requirements

Originally, there are 14 climate models worldwide, but for this project only 7 of them are going to be used due to the long execution time. The data is stored in standard format that is used by climate scientists. Additionally, we are going to analyse Europe area only. It is required to use big data techniques to reduce the processing time to less than 2 hours. These techniques will include: sub-spacing, parallel processing, display of the big-data. The flowchart depicts the core idea of program flow and expected working principles. All the major steps done, will be recorded in the logbook in the excel format.

## Related documents

|  |  |  |
| --- | --- | --- |
| **Component** | **Name (with link to the document)** | **Description** |
| Code base | <https://github.coventry.ac.uk/barkausa/big-data> | Program code for Matlab excluding the big data files |
| DDC report | <https://ieeexplore.ieee.org/document/6930157> | Find out more about the DDC algorithm from its authors |
| Flowchart | /flowchart.drawio and /images/flowchart.png | Find the program flowchart as a project file or image |
| Installation guide | /README.md | Contains short description from coding perspective |

## Terms/Acronyms and Definitions

|  |  |  |
| --- | --- | --- |
| **Term/Acronym** | **Definition** | **Description** |
| observation models |  | Model is a measurement made by a different organization, which is not completely accurate |
| SME | Simple mean ensemble | An average of all provided models (in our case 7) for every location point within Europe |
| DDC algorithm | Data density clustering algorithm | An algorithm that clusters data points. Its output is being used to create a CBE |
| CBE | Cluster based ensemble | More accurate ensemble then SME, where generation of it is much more complicated than SME |
| Sub-spacing |  | Dividing a current data set to smaller data chunks |
| Parallel processing |  | Usually, splitting independent tasks to run on a different thread to execute faster. Closely related to sub-spacing. |
| flowchart |  | Depicts the main logical program steps visualized using a special symbol syntax |
| logbook |  | An excel sheet, were the major project steps are being recorded |

# Risks and Assumptions

*<For assumptions consider e.g. maximum and minimum screen resolutions if displaying data, data file format, data size, hardware specifications.*

*For risks consider, e.g. file corruption, data file changes in future, bugs in standard libraries, or algorithm code not written by you.>*

# Out of Scope

*<Describe anything that you specifically do not want to achieve, that may be part of the overall project.>*

# System/ Solution Overview

*<Provide a short description of the software and solution being specified and its purpose.>*

# Context Diagram/ Interface Diagram/ Data Flow Diagram, Application Screen Flow, Sitemap, Process Flow

*<Provide any appropriate graphical representations that are relevant to the system and project such as a context/interface/data flow diagram, application screen flow, site map, or process flow. Add as many as needed. Note: This will be a simple diagram at the start of your project, it is not a full code function and data flow diagram.>*

# Project Management

*<Describe how you will manage your project, consider things such as small steps for weekly achievement, planning around other aspects of your course, e.g. exams, VIVA etc. and don’t forget version control of your code! >*

# References

*List all references to external material used as background information or knowledge.>*

# Open Issues

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Issue ID** | **Issue** | **Raised By** | **Raised On** | **Solution/ Decision** | **Resolved By** | **Resolved On** | **Status** |
|  |  |  |  |  |  |  |  |

# Appendices