D5.1 - User-driven Requirements & Use Cases

Co-funded by the Horizon 2020
Framework Programme of the European Union

DELIVERABLE NUMBER	D5.1
DELIVERABLE TITLE	User-driven Requirements & Use Cases - Agro-climatic and economic modelling
RESPONSIBLE AUTHOR	Rob Lokers (Wageningen Environmental Research)

GRANT AGREEMENT N.	731001
PROJECT ACRONYM	AGINFRA PLUS
PROJECT FULL NAME	Accelerating user-driven e-infrastructure innovation in Food & Agriculture
STARTING DATE (DUR.)	01/01/2017 (36 months)
ENDING DATE	31/12/2019
PROJECT WEBSITE	http://www.plus.aginfra.eu
COORDINATOR	Nikos Manouselis
ADDRESS	110 Pentelis Str., Marousi GR15126, Greece
REPLY TO	nikosm@agroknow.com
PHONE	+30 210 6897 905
EU PROJECT OFFICER	Mrs. Georgia Tzenou

WORKPACKAGE N. | TITLE	WP5 |Agro-climatic & Economic Modelling Community
WORKPACKAGE LEADER	Wageningen Environmental Research
DELIVERABLE N. | TITLE	D5.1 | User-driven Requirements & Use Cases - Agro-climatic and economic modelling
RESPONSIBLE AUTHOR	Rob Lokers (Wageningen Environmental Research
REPLY TO	rob.lokers@wur.nl
DOCUMENT URL	http://www.plus.aginfra.eu/sites/plus_deliverables/D5.1.pdf
DATE OF DELIVERY (CONTRACTUAL)	30 June 2017 (M6), 30 June 2018 (M18, 1st Updated version)
DATE OF DELIVERY (SUBMITTED)	17 July 2017 (M7), 29 June 2018 (M18, 1st Updated version)
VERSION | STATUS	V2.0 | Final
NATURE	R (Report)
DISSEMINATION LEVEL	PU(Public)
AUTHORS (PARTNER)	Rob Lokers (Wageningen Environmental Research), Rob Knapen, (Wageningen Environmental Research)
REVIEWERS	Matthias Filter (BfR)

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
1.0	Initial Word version	17-07-2017	Rob Lokers (Wageningen Environmental Research), Rob Knapen, (Wageningen Environmental Research)
2.0	1st Updated version	29-06-2018	Rob Lokers (Wageningen Environmental Research), Rob Knapen, (Wageningen Environmental Research)

PARTICIPANTS		CONTACT
Agro-Know IKE (Agroknow, Greece)		Nikos Manouselis Email: nikosm@agroknow.com
Stichting Wageningen Research (DLO, The Netherlands)		Rob Lokers Email: rob.lokers@wur.nl
Institut National de la Recherché Agronomique (INRA, France)		Pascal Neveu Email: pascal.neveu@inra.fr
Bundesinstitut für Risikobewertung (BFR, Germany)		Matthias Filter Email: matthias.filter@bfr.bund.de
Consiglio Nazionale Delle Richerche (CNR, Italy)		Leonardo Candela Email: leonardo.candela@isti.cnr.it
University of Athens (UoA, Greece)		George Kakaletris Email:gkakas@di.uoa.gr
Stichting EGI (EGI.eu, The Netherlands)		Tiziana Ferrari Email: tiziana.ferrari@egi.eu
Pensoft Publishers Ltd (PENSOFT, Bulgaria)		Lyubomir Penev Email: penev@pensoft.net

ACRONYMS LIST

GI	Geospatial Information
SDI	Spatial Data Infrastructure
OGC	Open Geospatial Consortium
VRE	Virtual Research Environment
GIS	Geographical Information System
NDVI	Normalized Difference Vegetation Index
GACS	Global Agricultural Concept Scheme
TSUM	Temperature Sum (cumulative temperature)
UAV	Unmanned Air Vehicle

EXECUTIVE SUMMARY

This document introduces the domain of agro-climatic and agro-economic modelling and its use cases to be implemented in the AGINFRA+ project. It defines the involved community and stakeholders and provides a set of typical "personas". Moreover, different use cases related to the identified personas are described, as well as the specific data, semantics and analytics and processing involved in the implementation.

The use cases described in this document focuses on opportunities to support the described personas and to bring them from a local, single-computer and mostly peer network based work environment to a cluster compute - cloud based (VRE-like) collaborative work environment. We have therefore described specific work processes that in most cases currently have not made that transition, but are suitable to be transformed provided that VREs can fulfil its requirements.

In the next phase of the AGINFRA+ project, selected elements of the described use case will be further specified (both process and technical), adapted and deployed with the support of AGINFRA+ partners and evaluated by the involved community of stakeholders.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 SCOPE

This document introduces the domain of agro-climatic and agro-economic modelling and its use cases to be implemented in the AGINFRA+ project. It defines the involved community and stakeholders and provides a set of typical “personas”. Moreover, different use cases related to the identified personas are described, as well as the specific data, semantics and analytics and processing involved in the implementation.

Use cases described in this document focuses on opportunities to bring researchers from a local, single-computer and mostly peer network based work environment to a cluster compute - cloud based (VRE-like) collaborative work environment. We have therefore described specific work processes that in most cases currently have not made that transition, but are suitable to be transformed provided that VREs can fulfil its requirements.

1.2 TARGETED COMMUNITY

The agro-climatic and agro-economic modelling community aims to assess challenges like food security, food safety and climate change impacts in an integrated manner. The mission of this research community lies in improving historical analysis and short and long-term forecasts of agricultural production and its effects on food production and economy under dynamic and multi-variable climate change conditions, aggregating extremely large and heterogeneous observations and dynamic streams of agricultural, economical, eco-physiological, and weather data. The community working on the implementation of such use cases is a diverse network of agricultural, climate and economic researcher, practitioners and service providers in the science, policy and business domains. This network supports a diverse group of stakeholders consisting of public and private policy and decision makers in the broad area of agriculture, food security, climate change adaptation and disaster risk reduction and their supporting staff. These stakeholders operate on all geographical levels from local to global scale and include among others governments, business, NGO’s, and a diverse range of interest groups.

2 AGINFRA+ USE CASE CHARACTERISTICS

2.1 INTRODUCTION

The use cases described in this report, and foreseen to be (partly) implemented in the AGINFRA+ VRE are covering the prediction of short term and long term agricultural yields in the face of both operational, within season decision making, and longer-term decision and policy making with respect to climate smart agriculture and climate adaptation.

A typical use case in this area would covers the following tasks:

- Data acquisition
- Data analytics
- Data (pre-)processing
- Modelling
- Data visualisation
- Data storage and data curation

Typically, data acquisition is a task that is performed in the initializing phase of the use case implementation, while analytics, processing and visualisation are relevant both before the modelling phase (pre-processing) and after the modelling phase (post-processing).
A large part of the data being processed in this working area is geographic information (GI) and time series, often combined, so spatio-temporal datasets. GI comes in both vector (e.g. shape files, well-known-text (wkt), etc.) and raster (grids, coverages, geotiff, netcdf, etc.) formats. Besides as files the data is also commonly stored in databases with specific support for spatial information storage and indexing (e.g. PostGIS, Oracle Spatial). The special nature of GI should not be underestimated, typically it is processed with specific applications (Geographic Information Systems) and over the years many international standards have been developed for exchanging and working with this type of data. The Open Geospatial Consortium (OGC, www.opengeospatial.org) and W3C (www.w3.org) maintain most of these standards which are widely used by the GeoScience community. Many countries have developed national Spatial Data Infrastructures (SDI) to support the sharing of and working with GI. Consequently many "legacy" formats and systems exist. Moreover, it is good to note that consequently the spatial domain takes a more closed world assumption to the data than the semantic web does, thus integrating them is not straightforward

2.2 DATA ACQUISITION

Data acquisition is an important stage in any modelling exercise, as it co-defines the baseline and is determinative for the quality of end results. In this use case, generally, the aim is to be able to discover, access and download for further usage a range of agro-climatic data resources that are relevant in the perspective of the targeted working processes. Data acquisition tasks thus involves the acquisition of data on weather and climate, crops and crop experiments, soil, land use.

2.3 DATA ANALYTICS, PROCESSING AND MODELLING

Preparing a modelling exercise usually requires a substantial effort with regard to data pre-processing and data integration activities. These vary from quality control (e.g. outlier and anomaly detection) and quality improvement (e.g. outlier and anomaly corrections, gap analysis and filling) to data harmonisation (e.g. re-scaling) and data integration (e.g. translation into modelling data schemata).

We foresee that dealing with such processing tasks will require substantial work on the area of data analytics and data processing with relatively large and heterogeneous datasets. Consequently, this might require the setup of a suitable big data processing and analytics environment for (explorative and

interactive) modelling, as a foundation for further work in the use cases. Care should be taken to also add the specific spatial operations usually performed by GIS to the processing environment. An example of this is region connection calculus (RCC8, https://en.wikipedia.org/wiki/Region_connection_calculus).
The actual modelling can be CPU intensive and can thus benefit from the grid computing facilities of a VRE. It should be noted that this requires models to be (made) fit for running on a grid, and might also have consequences for the pre-processing and post-processing of the data.

2.4 DATA VISUALISATION

Interactive data visualisation is an important asset in all steps of a modelling exercise. It is required to be able to explore input data, to examine results of (pre-)processing and data analytics tasks and to visualize and disseminate modelling results to end users.

Typically, agro-climatic and economic modelling is a spatio-temporal activity, so visualisations should support working along spatial and temporal dimensions (e.g. geographic maps and time series represented in graphs, maps and animations). Moreover, dealing with uncertainty (probability) is an important aspect of any modelling domain. As both short-term operational and long-term strategic assessments rely on weather and climate projections, which are uncertain by definition, the inclusion of uncertainty/probability in visualisations is even more important in this specific type of modelling.

3 PERSONAS

3.1 PERSONA 1 - RESEARCHER

The researcher's work is focussed at scientific and applied scientific definition and application of models. These could either aim at developing and applying innovative models and model applications or at configuring and applying existing models for innovative research on societal challenges like food security or climate change.

Alain Duvall - Senior Agronomic Researcher

"*I would like to find ways for me and my team to work more efficient and to****make it easier to share and reuse data and analytics***"

Name:	Alain
Position:	senior-researcher at an agronomic research organisation
Age:	46
Education:	University, PhD
Location:	France
Archetype:	Researcher

Biography

Alain has a PhD in Agronomy. He has worked as a researcher in France and the US for an NGO and several research organisations and he's been involved in developing simulation models and performing impact assessments for years. Since a few years he is head of a small research team. Besides, he works on a strategy to better exploit big data and big data analytics in his domain.

Daily Tasks

- Operational management of a small team of researchers and ICT professionals specialized in agro-climatic modelling and impact assessment.
- Coaching his team members in their daily work.
- Setting out directions for future research in the agronomic domain, with a focus on big data and data analytics.
- Scientific publication.
- Data curation.

Motivations

- Doing research that is being used and has impact.
- Progressing his scientific career.

Goals

- To provide good quality scientific outputs
- To extend his team's data analytics capacities
- To produce results that are fit-for-use for policymakers

Frustrations:

- Why is it still so hard to find the data I need for my research?
- All my team members are developing their own data analytics, there's too much redundancy and too little reuse.
- I cannot easily publish my data in a citable manner.

3.2 PERSONA 2 - INFORMATION INTERMEDIARY / SERVICE PROVIDER

Information intermediaries and service providers usually aim to provide added value, like farm management advice, through the(re)use of existing data and (meta)models to support end users (for instance policy and decision makers or farmers). They aim at providing operational services, using trusted datasets and proven technologies.

Kenneth Brown - ICT developer

*"Despite all efforts it is still hard to **find trusted data and process it efficiently for modelling**"*

Name:	Kenneth
Position:	programmer and data-analyst at AgroTalk, a small ICT company
Age:	28
Education:	University, BSc
Location:	Uganda
Archetype:	Developer

Biography

Kenneth has finished his BSc in informatics in Uganda and has just joined AgroTalk. AgroTalk is a small ICT company that specializes in providing advice to farmers through among others mobile phone and smartphone apps. With his six colleagues Kenneth is developing a service that provides farm management advice to smallholders through SMS services and through a farm management support system that is used by extensionists in the field. They are providing crop advice using a crop growth model that uses among others weather data. In the future, they plan to also develop a smartphone app for this purpose.

Daily Tasks

- Co-developing a software system that performs automated daily crop model runs and turns the results into advice.
- Manual and semi-automatic analysis of data sources like weather data, crop development data, soil data etc., to provide advice to farmers.
- Liaising with two colleagues, who are agronomists and advise him on his data analysis work and on the implementation of agronomic algorithms for the new IT system.

Motivations

- Programming reliable, robust software for smallholders.
- Using new technologies, with a special interest in big data and data analytics.

Goals

- To deliver a high-quality ICT solution that has added value and can be charged for.

Frustrations:

- The data I need to turn agronomy knowledge into working solutions is hard to find and get, e.g. weather data is available, but often too expensive, and the data I can get is not always of good quality.
- Data is provided in so many formats that are not compatible and there's no software to easily process the data.

3.3 PERSONA 3 - BUSINESS ANALYST

John Jackson - Policy advisor and analyst

"I know there is scientific work on agricultural data analytics but I do not have access to scientific resources other than papers"

Name:	John
Position:	Data analyst at the Ghana Ministry of Agriculture
Age:	36
Education:	University, BSc
Location:	Ghana
Archetype:	Data Scientist

Biography

John has been working at the Ministry of Agriculture for 12 years now. He started as a policy officer, but he has always had interest in data analysis and three years ago he has been transferred to the new department for data analytics and data science. He is producing regional crop bulletins, using data collected in the different agricultural regions of Ghana and data provided through the national weather service. The results of his work are published through the open data platform that the government has just set up. Currently he is exploring the opportunities of satellite derived NDVI data that his department has been provided access to recently.

Daily Tasks

- Collecting the data for his work from different sources.
- Analysing, improving and post-processing the crop and weather data as input for his data analysis algorithms and specifically exploring the options of NDVI data streams.
- Publication of his output data on the open data platform and curation of the section on agricultural data.

Motivations

- Working on this new field of data science, using new technologies and analytics.
- Promoting open data and the use of open data by farmers and business.

Goals

- Make his work easier by improving the data flows.
- Operationalize the analysis of NDVI time series
- Becoming a senior data scientist.

Frustrations:

- Data is very scattered and it takes too much effort to collect and integrate it.
- I do not have access to scientific resources for my data analytics.
- Even though we are providing open data, I see that it is hardly used.

4 DATA AND RELEVANT SEMANTICS

This section describes some typical data sets that we expect to be relevant in the elaboration of the use cases in this domain of AGINFRA+. More specific definitions of datasets, their specifications and their specific use in user actions will be described in the section use cases of this document.

4.1 RELEVANT DATASETS

AgroDataCube
AgroDataCube is an initiative by Wageningen UR to bring together a baseline of open data relevant to agro-climatic research for the Netherlands. Currently it is implemented as a PostgreSQL / PostGIS database containing weather, soil, land use and agronomic data. The system is a work in progress, including a roadmap to develop alternative ways to access the data other than using (spatial) SQL.

Global Soil Information Facility
ISRIC's global coverage soil data repository[1] contains global soil types at 1 km grid resolution and is freely accessible. Soil types are statistically derived and quality varies per country. Raw data can be downloaded from an FTP download service[2] (available in GeoTIFF format). Besides, it can also be accessed via a Web Coverage Service (WCS) and through a REST API. For more information see their website.

European Data Infrastructure
Currently the data available at the European data e-infrastructures (EUDAT, OpenAIRE etc.) is not specifically useful for agro-climatic modelling However, any agriculture and climate dataset as provided through this data infrastructure should be easily accessible and reusable for implemented VREs for this domain.

Open Data Journals
Over time, we expect that datasets published through open data journals can be used for modelling and vice versa end results can be added to into such journals. Most relevant on the short term would be the Open Data for Agricultural Research[3] (ODJAR), a citable open data journal for the agricultural domain. At the moment ODJAR contains only a few datasets and they can be downloaded as csv files. More data and other formats are expected in the near future.

ISIMIP data repositories
ISIMIP is the Inter-Sectoral Impact Model Intercomparison Project[4] which publishes large collections of model input and output data for agro-climatic research on a global scale. The ISIMIP initiative hosts a global repository of modelling data, containing among others future climate data projections and outputs of agronomic models. Typically, data is available in NetCDF format (using the Climate and Forecast conventions[5]) and can be downloaded from their servers although access is required. See their website for instructions.

NDVI data from the Dutch "groenmonitor"

[1] https://soilgrids.org/

[2] ftp://ftp.soilgrids.org/data/recent

[3] www.odjar.org

[4] https://www.isimip.org

[5] http://cfconventions.org

The "groenmonitor"[6] publishes NDVI data for the Netherlands derived from open satellite data repositories (mind that the website is in Dutch). NDVI is an often-used indicator for estimating crop development and is therefore often used as an alternative data resource in crop modelling and yield forecasting. Relevant data will be added to the already mentioned AgroDataCube and accessible from there.

4.2 SEMANTICS

Although efforts are ongoing to establish shared semantics for agro-climatic research, for example in the AgriSemantics initiative[7], this has still a way to go before it can be efficiently used in agro-climatic modelling use cases. The domain has developed from a range of initiatives that have their origin in the 20th century, with attempts on integration only started in the last couple of years. Consequently, many datasets will have their own naming conventions, terminology and ways for providing metadata.

AgroVOC[8] is a thesaurus that is commonly referred to when datasets are available or transformed into linked data. Currently an international task force of FAO, NAL and CABI is working on a self-funded initiative to create the Global Agricultural Concept Scheme[9] (GACS), identifying a set of concepts common to their three thesauri. When variables are harmonized across datasets (in particular for agricultural field experiment data the ICASA data standards[10] are often used.

[6] http://www.groenmonitor.nl

[7] http://www.agrisemantics.org

[8] http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus

[9] http://agrisemantics.org/gacs/

[10] http://www.sciencedirect.com/science/article/pii/S016816991300077X

5 DATA ANALYTICS AND DATA PROCESSING

Although impossible to be complete, this section describes a few of the typical data processing and analytics tasks performed in the field of agro-climatic research that are feasible for execution on a VRE. It should be noted that researchers involved might be used to performing all work on their desktop computer or even their laptop, working on local copies of data, and only have limited IT and programming skills (not their core expertise). They might be used to downscale the problem and data so that it fits their (more traditional) working environment, instead of learning the new skills required to be able to upscale to larger infrastructures. Furthermore, storage of data 'in the cloud' (on a Virtual Research Environment) triggers questions about data security and privacy. Datasets being worked on and results might be of a confidential nature.

Derivation of spatial relationships within agro-climatic big data

Derivation of (spatial) relationships between the core datasets available in the AgroDataCube, as a basis for both data processing and data analytics. Possibly also with some of the other datasets described in the previous section. Examples are the linkage of (administrative) agricultural parcels over different administrative years or the determination of meteo-stations nearest to a parcel through geo-spatial analytics such as overlays, distance calculations, zonal statistics.

Spatial data interpolations

The development and application of interpolation and scaling procedures to e.g. downscale station weather datasets to the agricultural parcel level or to harmonize data captured at different spatial scales and with different resolutions (e.g. harmonizing weather and soil data to a unified grid). A more detailed explanation of types of spatial analysis that are commonly performed by researchers can for example be found in the manual for QGIS[11] (a well-known open source GIS):

Mapping crop phenology to NDVI time-series

NDVI (Normalized Difference Vegetation Index) data provides information on a time specific state of the development of a crop and is derived from sensing images (remote or close sensing, e.g. satellite, UAV or farm machinery imagery). In the case of satellite imagery, which is the basis for the previously mentioned "groenmonitor", the amount of available and usable satellite data over time is restricted because of limited sampling frequencies and frequent disturbance through cloud coverage and cloud shading. As crop development can be characterized through sigmoid curves, fitting such a function to the available data could provide a more continuous picture of crop development and help fill in the gaps. Much research is being done on this mapping of crop phenology to the NDVI (or related indexes) time-series. For an example see this publication: https://doi.org/10.1016/j.jag.2014.08.011.

Derivation of aggregated climate and agricultural indicators required for crop modelling

In crop modelling and yield forecasting, aggregated indicators play an important role. For instance, temperature is an important driver for crop development. Derived from temperature series, as an example of a relatively simple aggregated indicator, cumulative temperature (TSUM) over the season is a relevant input for crop models. Means, averages, and moving averages are also frequently used statistics.

Determination of temporal and spatial correlations and patterns

Spatio-temporal analytics can reveal patterns that are relevant to, for instance, explain observed crop development anomalies. As an example, low yields could be explained from local weather conditions at crucial time periods in the development of a crop, e.g. at the time of germination. At the moment

[11] https://docs.qgis.org/2.2/en/docs/gentle_gis_introduction/spatial_analysis_interpolation.html

researchers use dedicated systems and visual inspection of data for this type of analytics (e.g. https://ec.europa.eu/jrc/en/mars). New methods such as Machine Learning could play a role and be beneficial to their work, in particular when working at more detailed scale and when amounts of data to process increase, but this is largely an unexplored territory for agro-climatic researchers at the moment.

6 VISUALISATION

Visualisation of datasets is required to be able to explore input data, to examine results of pre- or post-processing and data analytics tasks and to visually explain and disseminate modelling results to end users like policy and decision makers. Visualisations must support working along spatial and temporal dimensions (e.g. geographic maps and time series represented in graphs, maps and animations) and dealing with uncertainty (probability).

(Interactive) maps

As most data related to the agro-climatic domain is spatial, geographic maps are a common way to visualize data and the results of data analytics, processing and modelling tasks. Typically, interactive maps are used to explore (select layers, zoom, pan), compare (e.g. compare over time) and determine local values (e.g. crop yields at a specific point in the map). Some examples:

- Crop development indices, yield maps.
- Soil Maps.
- Weather data and derived weather data (TSUM, evaporation) maps.

Animated maps are sometimes used to get a better perspective on the spatiotemporal dynamics of simulations.

Graphs

Another frequently used type of visualisation are graphs. Line graphs are most common for showing the development of indicators over time. Bar charts can be used to compare over spatial units, over time periods, over scenarios etc. Network graphs are typically used to study and follow semantic relations in linked data. For example:

- Crop growth sigmoids (line chart).
- Comparison of values over spatial units (e.g. bar chart).
- Ensemble results, including uncertainty (e.g. line chart with average and bandwidth).

7 USE CASES

This section specifies the directions for implementation of a set of use cases that are typical for the personas described in section 3. The use cases, or specific elements in the use case, are all appropriate to be performed on a VRE. The use case will be further specified in the next phase of the AGINFRA+ project. Work processes and elements will be technically elaborated, data will be collected and components will be prepared for deployment on the VRE with support of the technical partners. In that phase, also the final decision on which (parts) of these use cases will be taken up by the project will be taken.

The following paragraphs provide a short description of the use cases from which elements will be selected. Table 1 describes typical tasks in the categories data discovery and caching, data processing, model execution and visualisation that could be executed on a VRE.

7.1 USE CASE - RESEARCHER

This use case focuses on the work of a scientific agronomic modeller and includes discovery and download of (raw) datasets, pre-processing of datasets (harmonisation, integration), running a crop growth model and analysis and visualisation of model outputs.

There are two sub cases that are feasible for VRE deployment and execution:

A. Regional yield forecasting
Finding correlations between historical yield (statistics) and indicators derived from remote sensing and climate data time series, using the strongest correlations for yield forecasting. An example and description of the methodology can be found in this article: http://www.sciencedirect.com/science/article/pii/S0168192315000702. It can be extended by looking at other indicators, e.g. precipitation and temperature sums, and find correlations with the NDVI time series (or the fitted growth curves).

B. Large scale regional crop model simulations
Most crop models are point based simulation models. With the proper input data (crop type, soil, climate, field management) available the model can be run in parallel for many regional locations, several crops, and input variations, and results combined. This can now only be done in a limited way using traditional hardware or e.g. Amazon Cloud Services.

7.2 USE CASE - INFORMATION INTERMEDIARY / SERVICE PROVIDER

This use case aims to include the main tasks of an information intermediary or service provider.
It focuses on the finding and integrating of data, and providing services to retrieve it. Semantics and metadata (including provenance) plays an important role (the European Open Science Cloud initiative increases the focus on data stewardship).Much of the agro-climatic data is spatio-temporal in nature, and there has been a long history in the development of Geo Information Systems (GIS) and Spatial Data Infrastructures (SDI) for working with these complex and large datasets. However, the standards developed within Geo-Information Science are academic and heavy based on closed world assumptions and need to be bridged to lighter community web and semantics standards.

There are two sub cases that are feasible for VRE deployment and execution:

A. Finding and selecting spatial and temporal subsets of data
The typical ways of discovering Geo-Information (GI) is (i) by asking known experts, (ii) through an internal department (e.g. a ‘Geodesk’ service point), or (iii) through local and global metadata

repositories. The NationaalGeoregister (http://www.nationaalgeoregister.nl/), for example, contains records of all GI datasets available as open data in The Netherlands (in particular those produced by governmental agencies). Once a suitable dataset has been found it needs to be retrieved, and available file format options may vary (e.g. GML, GeoJSON, KML, Shapefiles, or web access points such as WFS, WMS, WCS (standards defined by the Open Geospatial Consortium)). Retrieving data often includes selecting only a specific part of the data, the region of interest, to avoid downloading large datasets, and re-projecting the data so that in the end all datasets share a common spatial projection. These steps often result in a laborious manual process, partly due to (spatial) data quality issues needing to be solved. A VRE with a common data catalogue and data processing facilities with data provenance could help to reduce time spend on this data finding and selecting.

B. Spatial Data Wrangling
Just like with regular statistical data, integrating spatial data and shaping the data into something usable involves an amount of (spatial) data wrangling. Re-projecting datasets to make sure all data uses the same geographic projection is an example. Transforming between vector and raster data is a very common process as well, which might include upscaling or downscaling data to deal with differences in spatial (and temporal) resolutions. Cropping data to areas or cut-out polygons (e.g. country boundaries) to select study regions helps to reduce the amount of data before processing.Calculating zonal statistics, spatial overlays, centroid points for polygons, and distances between e.g. points and polygons, are as well amongst the most common initial processing steps to create a usable dataset. The VRE could help to run this kind of processing as repeatable workflows that can be stored for later re-use, and that can be executed on a cluster to reduce total processing time.

7.3 USE CASE – BUSINESS ANALYST

This use case focuses on data analytics, covering relevant steps of solving a data science problem. The case includes finding the data, wrangling it, exploration of results (e.g. with Python Pandas or RStudio), modelling, processing, visualising, collaboration/publishing.

There are two sub cases that are feasible for VRE deployment and execution:

A. Estimating growth curves
Remote sensing imagery, e.g. from satellites or drones, can be used to calculate variables such as the Normalized Difference Vegetation Index (NDVI). Typically, this gives an irregular time series of weekly values. By fitting a curve (e.g. a double sigmoid curve) through these data points an estimation can be made of crop phenology and hence expected yields. However, such curve fitting is not trivial and can be time consuming. E.g. http://www2.geog.ucl.ac.uk/~plewis/geogg124/phenology.html gives an example of the process. Using VRE compute capabilities calculating NDVI for large areas can be done in shorter time, and perhaps refined with machine learning algorithms that take more data into account.

B. Crop yield risk assessment - Detrending models
Detrending is a widely used technique for obtaining stationary time series data in residual analysis and risk assessment. The technique is frequently applied in crop yield risk assessment and insurance ratings. E.g. see http://link.springer.com/article/10.1007/s00477-014-0871-x. Since this is a very common practice for researchers and analysts it would be good if the algorithms required are supported by the VRE.

	Data discovery and 'caching'	Pre-processing (Focus for this project)	Model execution (Focus for this project)	Visualisation (Focus for this project?)
1st use case: Regional yield forecasting OR Large scale global crop model simulations.	Regional yield forecasting can be based upon the content of the AgroDataCube. Global simulation would need global data (climate scenarios?), soil. Crop data perhaps can be extracted from our internal database (though that is not open data). The use case would be to see if the infrastructure is usable for a system like CGMS (see info below).	Deriving data for the model (data preparation) Is this an interactive process and a 'batch job'. GIS environment is used today (e.g. QGIS). Usually the significant unique combinations of climate zones, soil, etc. are calculated in order to reduce total required model runs.	The model is WOFOST. This is a point based crop simulation model. Many implementations exist. The original model is available as Fortran source code. There is also an open source Python version available on GitHub (PCSE), and we are working on a Java implementation. UC Davis has a version in R (RWOFOST, we can try to get the package from them). WOFOST typically reads ***input from SQL databases or text files***, and writes **output into csv text files**. The simulation can be run independently for a single point, so multiple simulations can be run in parallel. Algorithms used in the model are very traditional and not parallelized.	A list of summary variables is given below. For a large scale simulation a spatial map of max. leaf area or total biomass would be a good starting point. From there a researcher would like to drill down into to the data for specific areas and see time series (charts) of the variables. And even further down to see time series of all the other (more intermediate) model variables and inputs.
2nd use case: AgroDataCube	AgroDataCube is an (internal) SQL (Postgresql) database that	Data in the AgroDataCube has no well-defined data	There is no predefined model. Examples of performed ***spatial***	Depends on the spatial query, usually the result is a

	Data discovery and 'caching'	Pre-processing (Focus for this project)	Model execution (Focus for this project)	Visualisation (Focus for this project?)
	contains a number of open datasets relevant to agronomy (daily weather, soil, yearly crop registrations, altitude, NDVI time series, administrative boundaries) for multiple years (2012-2016). All these datasets are stored as spatial vector or raster layers (tables in the database). The database uses the PostGIS extension to deal with spatial data and queries. Metadata and documentation is available for the open data sources used, but not duplicated into the database.	model. Extracting data is best done using spatial queries (e.g. geometric overlaps, zonal statistics, distance calculations, point-in-polygon, etc.). All data in the AgroDataCube has been re-projected to the Dutch RD coordinate system, so coordinate transformations from/to latitude,longitude might be needed too.	***analytics*** are: (1) calculate distance between all parcels of crop registration and all points (meteo-stations) of daily weather, find the nearest and interpolate weather data. (2) find overlapping parcels of crop registration from 2012 - 2016 to detect crop rotations in areas. Many more operations can be thought of. Calculations are highly parallelizable. Spatial indexing can improve performance significantly.	***spatial map*** (vector or raster), or a time series, that researchers directly view in a GIS (e.g. QGIS or ArcGIS).
3rd use case: Crop phenology estimation	Crop phenology estimation (crop growth curves) typically is done based on remote sensing data. Either satellite or aerial based data. For the Netherlands we have pre-processed satellite data (Groenmonitor), and global data is available from MODIS (NASA) and Sentinel (ESA).	Raw satellite imagery needs pre-processing to make it usable (need to ask experts for further details). E.g. selection based on cloud cover, geo-rectification, and image mosaicking.	From a good set of images data of interest can be derived from the spectral bands in the image. E.g. NDVI (normalized difference vegetation index). Using the available NDVI data points a growth curve can be fitted through them (e.g. a double sigmoid curve). Based on the curve information about crop phenology can be	Data is usually displayed as ***spatial maps*** for a specific date (e.g. calculated NDVI), or graphs for time series NVDI for a specific location, and for growth curves.

	Data discovery and 'caching'	Pre-processing (Focus for this project)	Model execution (Focus for this project)	Visualisation (Focus for this project?)
			derived. This data analytics work is usually done in an interactive environment at first (**R** and **Python** / Jupyteror Zeppelin notebooks) and later automated for bulk processing. E.g. python script run in parallel for multiple locations, or transformed into ***Spark***.	

Table 1: Use cases and typical tasks