Impact of Similarity Metric Selection for Multiclass Scenario Discovery of Land-Use Change Models

Master thesis submitted to Delft University of Technology

in partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE

in **Engineering and Policy Analysis**

Faculty of Technology, Policy and Management

by

Omar Quispel

Student number: 4107950

To be defended in public on March 11th 2021

Graduation Committee

Dr. ir.	J.H. Kwakkel	Section Policy Analysis
Dr.	T. C. Comes	Section Transport and Logistics
Dr.	N.Y. Aydin	Section Systems Engineering

Executive Summary

Land-use change models are a tool used for understanding land-use change and providing advice to policy makers. Since land-use dynamics are difficult to capture, and often contain uncertain aspects, incorporating this uncertainty is paramount. Currently, land-use change models follow a story driven approach where different scenario narratives are designed by stakeholders, before they are simulated using the model. However, this approach runs the risk of overlooking outcomes of interest as the designed scenarios might not account for all relevant outcomes.

To remedy this, scenario discovery can be used in tandem with land-use change modelling to cover the full uncertainty space. Scenario discovery uses large sets of model output to first group outcomes with similar characteristics together by assigning a threshold value for desired and undesired outcomes. Afterwards, input parameters are then analyzed to find the driving forces that lead to a given set of outcomes. Even so, applying this approach to land-use change models requires a different approach than usual.

Because the output of land-use change models is multiclass in nature, a multiclass classification of outcomes is necessary in favor of a binary threshold value usually encountered with scenario discovery. Due to this multiclass nature, similarity metrics are used to compare the maps generated by land-use change models. A new step that is also required, is that the outcomes of the map comparison using similarity metrics are clustered based on their similarity. The resulting cluster data can then be used as input for the rule induction algorithm which tries to identify regions in the uncertainty space with a high concentration of results of interest. One of the key challenges with this procedure lies in the use of similarity metrics to compare maps.

It is currently unclear what the effect is of a given similarity metric on the scenario discovery process. To pave the way for combining scenario discovery and land-use change modelling, this research attempts to answer the question:

Main Research Question

How does the choice of similarity metric affect scenario discovery results of land-use change modelling?

Identifying a suitable set of metrics was done through the use of a literature review. This led to a total of ten similarity metrics, featuring kappa and Pontius' components of difference as a counterpart to kappa. Some of these metrics are reused to study different land-use categories. With these metrics, next the infrastructure was created to compare maps on a large scale. Some metrics were calculated using the Map Comparison Kit, while the rest were calculated directly in Python.

The data set required for this research was generated by the Land Use Scanner model, which has seen seen numerous iterations and has been applied in a policy context multiple times. With the model output processed and compared, the resulting confusion matrices were used as a basis for clustering. After performing an initial analysis of the clustering results, the clustering results were used as input for the CART analysis. Classification and Regression Tree (CART) analysis is a decision tree learning technique that forms decision trees based on variable levels of the data. It provides insight into which

input variables of the model played a big role in its results. Following this approach led to the following key findings:

Key Findings

- Of the 91 metric pairs compared, only 6 showed similar or the same results
- Metrics with a different mathematical approach can still lead to the same or similar results (i.e. kappa and total difference)
- Similarity in cluster results may still lead to noticeable differences in later scenario discovery steps
- It seems plausible that similarity metrics can be classified as being usable for a specific purpose
- If there is a lack of outcome variance (information) in land-use change modelling results, metrics may provide similar results
- Normalizing a metric does not necessarily provide the same results as its non-normalized version
- A noticeable difference was observed between similarity metrics in the use of input variables for their respective CART trees

This research has shown that using different similarity metrics to compare maps in the scenario discovery process is likely to lead to different outcomes in later steps. However, more work is required before scenario discovery can be used in tandem with land-use change models to provide policy advice. While this research showed that different similarity metrics generally lead to different clusters of maps, it is unclear what the different clusters of a given similarity metric represent. It is recommended that further research looks at combining scenario discovery and land-use change modelling with a wider set of models and similarity metrics. The purpose here should not just be to test more similarity metrics, but also to reuse metrics in different contexts and compare new and old results.

Preface

This report was written as part of the graduation process for the Master of Science program Engineering and Policy Analysis at Delft University of Technology. The chosen topic was an extension to the work of a PhD student working at the Faculty of Technology, Policy, and Management.

Most of the analysis in this research was processed using Python. While the main text focuses on the results, the appendices elaborate and explain the different Python files used and what they accomplish.

First I would like to thank my supervisors Jan Kwakkel and Bramka Arga Jafino for their guidance in getting this to where it is, but mostly for helping me get excited about getting the work where it needs to go.

Second, thanks go out to Eric Koomen and Jip Claassens from VU Amsterdam for providing access to the Land Use Scanner model and helping us get started. Special thanks to Jip for helping with the special requests we had to get the model doing what we needed it to do.

Finally, I am grateful for the help Hedwig van Delden from the Research Institute of Knowledge Systems (RIKS) provided with understanding the usage of similarity metrics for comparing maps. She provided a new perspective on things. Also from RIKS, thanks to Roel Vanhout for providing assistance with the Map Comparison Kit.

Omar Quispel
Delft, January 2021

Contents

Chapter 1

Introduction

Land-use change models are often used to explore future land-use and to help understand the effect of factors that affect land-use (Dalla-Nora, de Aguiar, Lapola, & Woltjer, 2014). Creating models that can do so reliably requires land-use change models that are meticulously calibrated and tested with empirical data. While this has proven to be a successful approach, land-use change models also have to accommodate for the inherent uncertainty of the factors that affect land-use change. Factors relating to demographic, economic or climate change aspects are inherently uncertain. Incorporating the uncertainty of factors in land-use change models is usually achieved by analyzing a small number of predetermined scenarios (see for example Rounsevell et al. (2006) and Ustaoglu, Williams, Petrov, Shahumyan, and Van Delden (2018)). To better address the multidimensional nature of uncertainty about the future, previous studies have argued for better handling of uncertainty (van Vliet et al., 2016; Verburg, Schot, Dijst, & Veldkamp, 2004).

Scenario discovery is an approach that implements this concept (Bryant & Lempert, 2010). The general approach followed in scenario discovery is that, instead of running the model a small number of times based on predetermined scenarios, the full spectrum of uncertainty is covered for the purpose of creating scenarios. The output generated by these scenarios is then classified as being of interest or not. This is usually achieved through the use of a binary classification on a variable that is considered important. If the goal of the research is to better understand how different land-use factors affect a policy objective, then this binary threshold could be as simple as a result satisfying the policy objective or not. The results that are considered of interest are then further analyzed with a rule induction algorithm. While all the remaining results are considered of interest, the rule induction algorithm provides a better understanding as to what variables lead to what kind of results. The benefit of this approach over using a small set of scenarios is that it is able to uncover relations between variables that can be overlooked when using a small number of scenarios.

But the use of a binary classification to classify policy success or failure does not suffice for land-use change models. For land-use change modeling, the purpose of analysis is about exploring plausible land-use patterns, and how these patterns come to be, rather than classifying model outcomes into those that are desirable and those that are not. Maps representing land-use maps are not simply the sum of all the land-use cells. Important information is instead often found in the spatial relationships between these elements (Hagen-Zanker, 2006). Simulation outputs can be better understood by clustering them based on similarity metrics that are already being used to compare maps. This leads to multiple clusters of interest that can be analyzed. Unlike in a binary classification approach, there is less chance of interesting outcomes being discarded on the basis of a binary threshold value. While such a binary threshold is easier to apply, it can hinder the model's ability to properly share the information that it contains. Additionally, land-use change models inherently contain multiple land-use categories which are represented in the maps that they output. Applying a binary classification to such outcomes would be challenging, and is unlikely to lead to satisfying results as information would undoubtedly be lost.

Classification of land-use maps is different from other modelling approaches used in tandem with scenario discovery because model output does not take the form of a set of variables that vary in importance to the users. Instead, maps are generated, after which the policy questions decide what kind of information needs to be obtained from these maps. In some cases this is similar to other modelling approaches in the sense that some similarity metrics provide a unique value to describe the characteristics of a map. In other cases, maps must first be compared to one another on a one-to-one basis to highlight their differences. Nonetheless, as classifying land-use maps is done through clustering, the former case still needs to compare the unique map value to all other maps generated by the model in question. This provides a contingency matrix for both cases that forms the basis for clustering the maps. Following this process allows for the separation of distinct model behaviour (Steinmann, Auping, & Kwakkel, 2020).

Therefore, this research focuses on understanding how different ways of classifying land use maps can be embedded within scenario discovery, and how this affects the quality of the scenario discovery results. Specifically, to move from a binary classification to a multiclass classification of model outcomes in the context of land-use change models, the identification of plausible distinctive land-use patterns through clustering algorithms is necessary. Before clustering can happen, the generated maps must first be compared amongst each other using similarity metrics. This is an additional step not usually required for traditional scenario discovery.

This research will provide initial insight into how the choice of similarity metrics affects the outcome of multiclass scenario discovery for land-use change models. While this has been explored before (Cox, 2020), this work focused on comparing story and simulation to scenario discovery results in the context of land-use change models. As such, only one similarity metric was used for the clustering of results. This research will focus on the effect that different similarity metrics have on scenario discovery results. From a methodological standpoint, the proposed research paves the way for further exploration of multiclass classification of results in favor of binary classifications in scenario discovery studies.

Chapter 2 starts by introducing the research questions, which have guided this research, before elaborating on how these questions were answered. Chapter 3 explains how scenario discovery is applied to land-use change models. Chapter 4 is the first analysis chapter and presents clustering results for individual metrics. Chapter 5 builds on this by then comparing clustering results between different similarity metrics, while also introducing classification of results. The research findings and its limitations are discussed in chapter 6. Finally, chapter 7 presents the conclusion to the research, as found by answering the research questions.

Chapter 2

Research Problem

This chapter elaborates on the research problem. First, the context of the research is presented, culminating in the research gaps. Following this, the research questions that guide the research are discussed. The overall structure of the research is highlighted afterwards. The choice of land-use model is then presented. Finally, the research contribution is covered.

2.1 Research Context

Creating a small set of scenarios to encompass the full scope of uncertainty is a challenging undertaking (Bryant & Lempert, 2010). Creating these scenarios can follow three schools of reasoning: Intuitive Logics Models, La Prospective Models, and Probabilistic Modified Trend Models (Bradfield, Wright, Burt, Cairns, & Van Der Heijden, 2005). With regards to land-use change modelling, the Intuitive Logics approach is generally used, in the form of a story and simulation approach. Story and simulation combines the best aspects of qualitative and quantitative scenario approaches allowing for both understandable narratives and the numerical data to back it up (Alcamo, 2008). Such an approach allows for a limited number of outcomes which are then easy to compare and process.

This research explores the full scope of uncertainty through the use of scenario discovery (Lempert, Bryant, & Bankes, 2008). Instead of creating a small set of scenarios before simulation, scenario discovery allows for a model driven approach. This is accomplished by running the model over a wide, to full, range of the uncertainty space that the model contains. Where in traditional scenario development the user has an ex ante role, in scenario discovery the modeller takes on an ex post role as the model is not limited to a small set of constraints. Because this approach generates much data to analyze, a different approach is required for analysis. Usually this takes the form of a binary classification where results are either of interest or they are not. Interesting results are then analyzed with a classification algorithm to find what input parameters lead to outcomes of interest.

There are however some key differences with how this research applies the scenario discovery method.

- land-use change models are the object of interest to be analyzed over the full spectrum of their uncertainty.
- multiclass scenario discovery is performed instead of binary classification of results

This leads to new challenges as land-use change models do not provide simple numerical output, instead they provide maps which all contain a wealth of information. As such, a binary classification of outcomes no longer suffices. Instead, maps have to be compared on a one-to-one basis before scenario discovery can continue. Additionally, excluding this binary classification means that there is no longer an arbitrary value that decides whether an outcome is of interest or not. Each outcome will be analyzed in later steps of the scenario discovery process, lowering the risk of discarding outcomes that can provide relevant results. This dichotomy between regular scenario discovery and the application of it in this research is summarized in Figure 2.1.

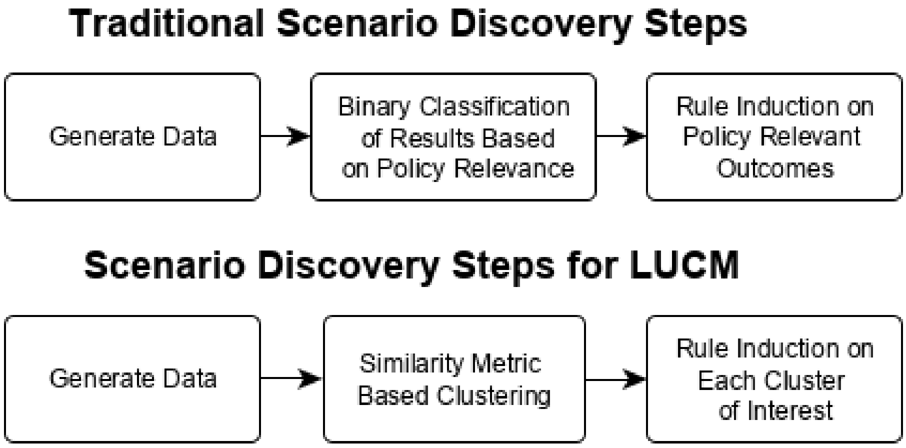

Traditional Scenario Discovery Steps
Generate Data
Binary Classification of Results Based on Policy Relevance
Rule Induction on Policy Relevant Outcomes
Scenario Discovery Steps for LUCM
Generate Data
Similarity Metric Based Clustering
Rule Induction on Each Cluster of Interest

Figure 2.1: Comparison of traditional scenario discovery steps

All discussed previously then leads to the following research gaps, which form the core of this research:

Research Gap 1

Little to no precedent of comparing land-use change model maps on a large scale.

Previous studies regarding the quantification of land-use map (dis)similarity focus only on comparing maps on a one-to-one basis or in small numbers for validation and calibration purposes (see for example Rounsevell et al. (2006), Ustaoglu et al. (2018)). While maps are sometimes run on a large-scale, this is generally limited to validation and calibration purposes and not for output analysis.

Research Gap 2

Noo precedent for how similarity metric selection affects clustering of land-use change model results.

In this study the effect of different similarity metrics on the resulting clusters of land-use patterns will be investigated systematically. Specifically, the implications of using various cell-by-cell similarity metrics as well as landscape structure metrics for clustering the land-use maps will be tested. While the work in this research will not be enough to provide a comprehensive framework on what similarity metrics to use for which policy questions using scenario discovery with land-use change models, it takes a first step towards such work.

Research Gap 3

Little to no precedent of land-use change model output being used for scenario discovery.

Research gap 3 builds on the information from research gap 1, allowing the model to dictate what information it has to share instead of the model builder dictating what is of interest. However, to do so, research gap 1 must be closed so that the process it develops can be used in further analysis steps. Research gap 3 then focuses on processing the cluster results further using machine learning algorithms, and analyzing these outcomes. This has been done before many times in the context of scenario discovery, but is relatively new for the case of multiclass classification of outcomes and only explored once with regards to land-use change model output (Cox, 2020). Exploring this path will allow land-use change models to be approached in a way never done before.

2.2 Research Questions

This research is guided by the research questions that are formulated based on the three research gaps discussed earlier. While the coming together of land-use change models and scenario discovery has numerous aspects that need exploration, this research tries to focus on a single one of the aspects. The main research question highlights the focus on the effect of different similarity metrics, while four sub questions are used to help answer the main research question.

Main Research Question

How does the choice of similarity metric affect Scenario Discovery results of land-use change modelling?

The main research question emphasizes the potential impact of different similarity metrics on scenario discovery results. Given available resources, this research cannot provide a full framework on the usage of similarity metrics when comparing maps. In part this is because, as far as the author knows, there is no precedent for this kind of work. Even more so, this research question will require testing many similarity metrics across various land-use change models. For this research, a satisfactory answer to the main research questions highlights initial differences when using different similarity metrics to cluster maps generated by land-use change models as part of the scenario discovery process. This paves the way for further exploration into the use of different similarity metrics.

Sub Research Question 1

What similarity metrics exist for comparing maps?

This question will be answered through a literature review. Articles that make use of similarity metrics for comparing maps will be scrutinized and compared. Based on this aspects that are likely to impact the scenario discovery process will be identified, and a subset of commonly used similarity metrics is selected for further use in the research.

Sub Research Question 2

How can similarity metrics be embedded in an agglomerative mode clustering algorithm?

The basic outline of how to embed the metrics will follow standard clustering procedure. However, it is conceivable that not all metrics identified in response to research question 1 can be fitted within this approach. The answer to this question is an algorithmic outline and associated proof of principle implementation in Python.

Sub Research Question 3

What is the effect of the different similarity metrics on the resulting clusters of maps?

This question will be answered using a case study. The metrics from research question 1 will be applied to a large collection of maps generated using the Land Use Scanner model, and the resulting clusters will be derived, described, and compared using confusion matrices and other tools for comparing clustering results.

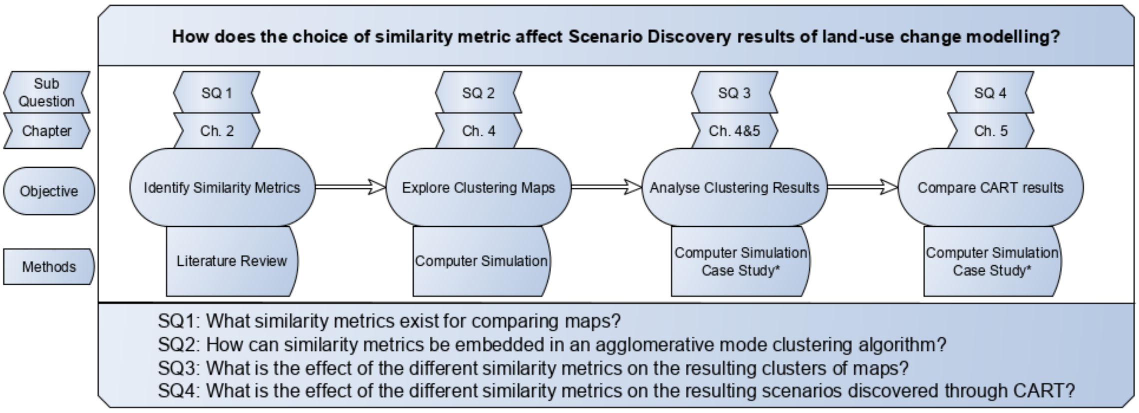
Sub Research Question 4

What is the effect of the different similarity metrics
on the resulting scenarios discovered through CART?

Building on research question 3, CART is applied to the resulting clusters to identify subspaces, or scenarios within scenario discovery parlance. CART is used over other options as it is the default algorithm to use with multiclass data (Gerst, Wang, & Borsuk, 2013). In this research the multiclass data consists of the generated maps and their land-use categories. This then allows for various aspects of the clustering process to be analyzed. One question that arises is whether the resulting subspaces are different across metrics. Of interest are also to what degree similarities in clustering results lead to similarities and differences following CART analysis.

2.3 Methods

The method used for answering each research question, and where it is covered in the report is summarized in Figure 2.2. The first objective is to better understand the usage of similarity metrics for comparing maps in land-use change modelling, and to select a set of similarity metrics for clustering. This is achieved through a literature review. Following this, the selected set of metrics must be embedded within the scenario discovery approach. A hands-on approach using computer simulation is used to identify whether map clusters formed based on similarity metric results can be used as input for CART analysis. The clustering results are then scrutinized in earnest using a case study and computer simulation. This step focuses on better understanding how each similarity metric leads to its formed clusters. Finally, the clustering results are used in CART analysis to see what impact the selection of similarity metrics has on the generated narratives. This is also achieved through computer simulation of the case study.



How does the choice of similarity metric affect Scenario Discovery results of land-use change modelling?

SQ1: What similarity metrics exist for comparing maps?
SQ2: How can similarity metrics be embedded in an agglomerative mode clustering algorithm?
SQ3: What is the effect of the different similarity metrics on the resulting clusters of maps?
SQ4: What is the effect of the different similarity metrics on the resulting scenarios discovered through CART?

* The case study is not used to study a specific situation, but to ground the analysis on a model that has seen validation and use.

Figure 2.2: Research Flow Diagram

2.4 Case Study Description

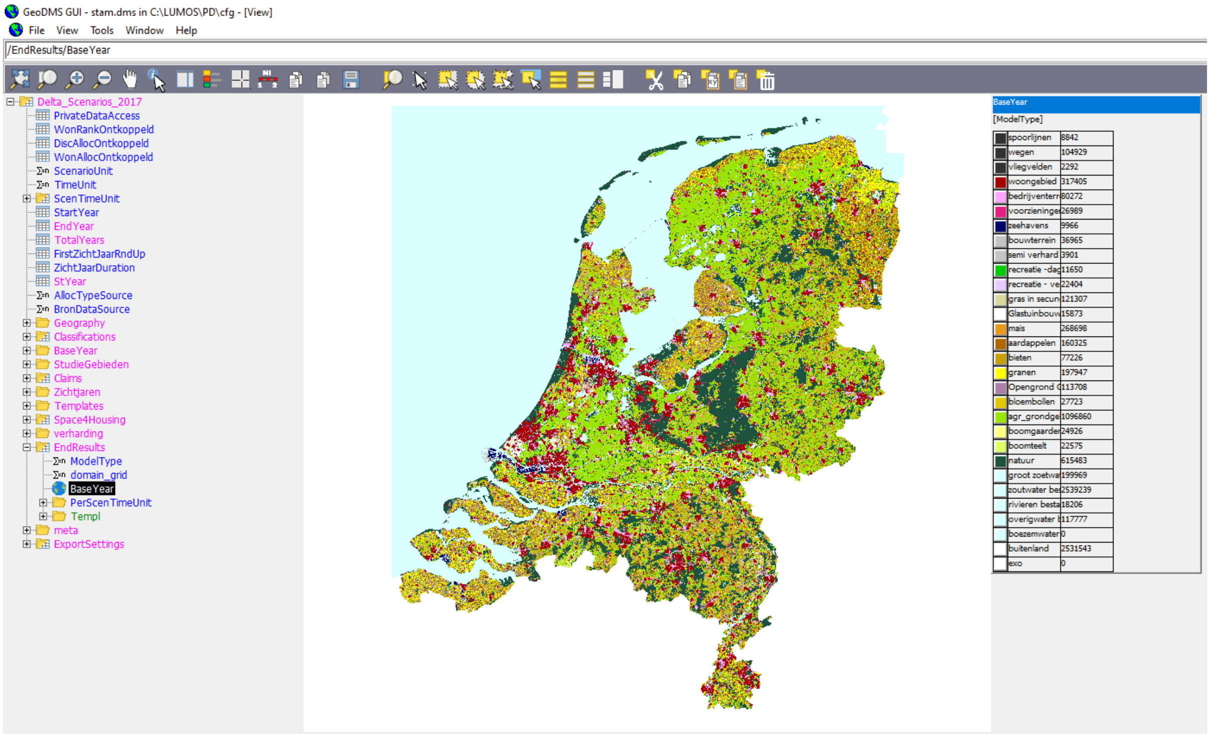
This research uses the Land Use Scanner model to generate maps for the scenario discovery process The usage of this model is not for the purpose of studying the case itself. Instead it is used because it is a tested and validated land-use change model that has seen numerous iterations and applications (Borsboom, Regt, & Schotten, 2002; Claassens, Koomen, & Rijken, 2017; Schotten, Goetgeluk, Hilferink, Rietveld, & Scholten, 2001). Any existing and validated model would have been a suitable option. Accessibility to the model thus played a large role.

The Land Use Scanner is a model created by Vrije Universiteit Amsterdam in close collaboration with the Netherlands Environmental Assessment Agency, Geodan, and the Agricultural Economics Research Institute. The authors describe the key features as being (SPINLab, n.d.):

- Grid based system that describes the relative proportions of land-use for each grid
- The model is able to accept multiple sector-specific databases as input for simulation
- The selection of land-use categories is exhaustive, nothing is left out
- The model is dynamic, adjusting itself to current land-use and being able to be run in multiple time steps
- Input data is based on national or regional forecasts
- The model is policy oriented as the sector specific data highlights policy conflicts

The model runs in the Geo Data and Model Software (GeoDMS) package. Figure 2.3 provides a snippet of LUS output in the GeoDMS interface, generated for what is defined as the 'base year' in the model with the allocated land-use in the table on the right of the image (in Dutch). The Land Use Scanner computes these maps on the basis of regional demand and local suitability factors. The regional demand follows from expert judgement or other models, which the Land Use Scanner uses as input. The local suitability follows from current land-use, suitability maps, or the inclusion of policy maps.

The model itself has no stochastic aspects. Given the same input, the output will always be the same. However, there are plenty of levers on the model to tweak. This research follows previous work where a set of uncertain factors for the Land Use Scanner were demarcated and explored (Cox, 2020). These uncertain factors pertain to economic and demographic developments, but also the impact of government action. The availability of, and encroachment on, nature is largely decided by policy maps.



⬛	spoorlijnen	8842
⬛	wegen	104929
⬛	vliegvelden	2292
🟥	woongebied	317405
🟪	bedrijventerr	80272
🟥	voorzieninger	26989
🟦	zeehavens	9966
⬜	bouwterrein	36965
⬜	semi verhard	3901
🟩	recreatie -dag	11650
🟪	recreatie - ve	22404
🟨	gras in secun	121307
⬜	Glastuinbouw	15873
🟧	mais	268698
🟫	aardappelen	160325
🟫	bieten	77226
🟨	granen	197947
🟪	Opengrond C	113708
🟨	bloembollen	27723
🟩	agr_grondgel	1096860
🟨	boomgaarder	24926
🟩	boomteelt	22575
🟩	natuur	615483
⬜	groot zoetwat	199969
⬜	zoutwater bes	2539239
⬜	rivieren besta	18206
⬜	overigwater b	117777
⬜	boezemwater	0
⬜	buitenland	2531543
⬜	exo	0

Figure 2.3: Land-Use Scanner example output for the 'Base Year'

Chapter 3

Scenario discovery for Land-Use Change Modelling

This chapter describes the process of applying scenario discovery to land-use change models. First, a short dive into multiclass scenario discovery is presented. Next the usage of similarity metrics in land-use change modelling is studied through a literature review. Following this, metric attributes that are relevant for clustering are presented. Based on preceding information, a set of similarity metrics is selected to be used as a basis for clustering land-use change modelling maps on a large scale. Afterwards the approach and results for finding the optimal cluster numbers are discussed. Then the analysis framework is shared. Finally, the experimental setup is presented.

3.1 Multiclass Scenario Discovery

Scenario discovery is becoming a widely used tool for scenario development (see for example Argent et al. (2016); Halim, Kwakkel, and Tavasszy (2016); McJeon et al. (2011). It can be loosely summarized as requiring the following steps (Steinmann et al., 2020):

1. **Generation** of simulation results
2. **Cluster analysis** of model results
3. **Rule Induction** on clusters to find what inputs lead to similar outcomes

The first step in the scenario discovery process is the generation of simulation results, or output data. This is achieved by running the model of choice across a wide, to full, range of the uncertainty space of its input parameters. In many modelling approaches only a small number of variables are changed, and then only for a few levels of each variable. For the purpose of scenario discovery, all model variables and their potential uncertainty are scrutinized. This creates an extensive set of scenarios that the model can be run under.

In multiclass scenario discovery these results are then clustered into groups that can be considered similar. For land-use change models, the distance value used for clustering is based on similarity metrics used for map comparisons. As each similarity metric has a specific mathematical approach to comparing maps, each cluster should consist of maps that are similar based on a similarity metric's mathematical focus. These clusters then form the basis for rule induction algorithms.

Algorithms used for rule induction aim to identify subspaces that contain the most outcomes of interest by using a bounding box. This is achieved by balancing three competing objectives (J. Kwakkel, 2019):

1. **Coverage** is the fraction of outcomes of interest contained within the subspace versus the total number of experiments of interest in the identified
2. **Density** provides the fraction of outcomes of interest relative to outcomes that are not of interest contained within the subspace
3. **Interpretability** refers to the number of restricted parameter space dimensions which limit the size of the bounding box

In this research, the rule induction step is of lesser importance than the the identification of model results of interest. In traditional scenario discovery this is done using a threshold value to identify under which conditions policy goals are not met (Bryant & Lempert, 2010). The downside to this approach occurs when the situation makes it challenging to agree on what this threshold should be. The nature of some models can also create situations where such a value would be less accurate or downright impossible to agree upon.

There is however precedent for using a multiclass classification approach instead of a binary one. Multiclass scenario discovery abandons the binary classification of results and instead clusters all model outcomes into different groups based on a distance metric of choice. This allows for all outcomes to be analyzed further. Rozenberg, Guivarch, Lempert, and Hallegatte (2014) still use threshold values, but they do so with multiple criteria at once. Gerst, Wang, and Borsuk (2013) use multiclass scenario discovery for an agent based model that focuses on economic growth, energy technology, and carbon emissions. Behaviour based scenario discovery applies a multiclass approach through the use of time series clustering (Steinmann et al., 2020). Only one example exists where scenario discovery was used in the context of land-use change modelling (Cox, 2020).

Because land-use change models inherently provide multiclass output, multiclass scenario discovery should be used over traditional scenario discovery. Unlike the previous two approaches, where either multiple criteria are used, or a single model variable is selected, multiclass scenario discovery for land-use change modelling must be performed by first creating confusion matrices for the maps based on similarity metrics, which are generally used to compare maps, before using these matrices as a basis for clustering.

3.2 Similarity Metrics for Land-Use Change Modelling

To obtain a better understanding of existing map comparison methods a literature review was performed. However, this is not a review of calibration and validation methods for land-use change models. Instead the objective is to find, and understand, how maps are compared, and whether this can be applied on a large scale to serve as a basis for clustering before applying a rule induction algorithm.

3.2.1 Similarity of Maps

Spatial analysis and the comparison of maps is relevant for a wide variety of topics, such as urban growth (Al-Shalabi, Billa, Pradhan, Mansor, & Al-sharif, 2012), ecology (Nagabhatla, Finlayson, & Senaratna Sellamuttu, 2012), water management (Tong, Sun, Ranatunga, He, & Yang, 2012), and risk mitigation (Nussbaumer, Huggel, Schaub, & Walz, 2013). Maps can be compared for a number of reasons, but six of them are commonly found (van Borsboom et al., 2004). These six reasons are:

1. Comparing model runs of different scenarios
2. Analyze temporal change in maps
3. For calibration and/or validation of land-use change models

4. Provide insight in why there are differences between maps
5. Identify so called problem areas that warrant further attention
6. Comparing different methodologies (i.e. different modeling approaches)

Unfortunately, the practise of map comparison, like many fields, suffers from a gap between the state of the art and state of the practise (Hagen-Zanker, 2006). As such, using previous application of map comparison methods simply because they have been applied before might lead to a continuation of this trend. Keeping up to date with methodological articles and exploring cross-disciplinary options is thus also of value.

3.2.2 Metric Redundancy

Over the last two decades there has been an increase in the number of metrics used (Tong and Feng, 2020). However, this increase in metric variety has also led to metric redundancy. As there are only a few primary measurements available for land-use change analysis, most metrics make use of these primary measurements in their own way (McGarigal, 2015). This can create redundancy through the addition of metrics that might sound different, but mathematically add little value. As such attention must be paid to what a metric calculates. However, as the goal is not to study only different metrics, some of this metric redundancy is in fact something this research will explore. To allow for this it is important to know what a metric is trying to accomplish and whether it aligns with the objective with which they are applied (Stehman, 1997).

3.2.3 Metric Exploration

A rough exploration of land-use change modelling literature was performed to better understand the usage of similarity metrics, and the degree to which they are used. While this review focuses on similarity metrics that are commonly used, issues highlighted earlier are still considered. Combining state of the art and state of the practise as mentioned in section 3.2.1 is attained by not only looking at metrics that see high usage, but also allowing for metrics that might not be as prominent for map comparison methods. The issue of metric redundancy discussed in section 3.2.2 is managed by exploring the mathematical background of any metric selected for further use. This is done in section 3.3.

The literature review of similarity metrics was largely performed on the basis of two exhaustive review articles that already provide the information that this research requires. In van Vliet et al. (2016) the authors compare the calibration and validation of models in articles published between 2010 and 2014. This time frame was chosen to compare recent examples as they were worried earlier articles would report on calibration and validation techniques that are less formalized. In a more recent article, Tong and Feng (2020) review calibration and validation techniques from 1999 to 2018. Contrary to van Vliet et al. (2016), they provide an overview over a longer time span. Of interest within these articles is that they summarize the usage of similarity metrics for map comparison across a large number of articles.

An overview of a large number of the metrics found in both articles is presented in Table 3.1. Due to the differences in presentation between both articles the information that could be extracted varies. While Tong and Feng (2020) provided a visual overview that was quick to interpret, a reference to the studied articles was lacking. On the other hand, van Vliet et al. (2016) provide schematic access to all literature cited through an excel file.

It is apparent that overall accuracy and kappa are most commonly used according to Tong and Feng (2020). As there is no reference to the articles, it is unclear whether this is because their timespan starts earlier, or whether this trend continues all throughout the observed timespan. It is possible that metrics such as allocation and quantity disagreement are always used together, explaining their equal usage. The absence of specific landscape metrics in van Vliet et al. (2016) is partially because they did not specify which landscape metrics were used, and grouped them all under the same tag.

The lack of the use of overall accuracy in van Vliet et al. (2016) could be explained partially due to simplistic, and thus unclear, tags used to indicate assessment methods. This could have occurred for other metrics, but seems most likely in this case.

The difference in metrics observed is likely in part due to the timespan, however there was also a difference in the search terms used and the search engine of choice. van Vliet et al. (2016) used model oriented terms in the web of science engine, whereas Tong and Feng (2020) used basic cellular automata and land-use terms in google scholar.

Side-by-side these two articles provided an overview of model assessment as practised in the land-use change modelling field. Combined, they highlight a variety of similarity metrics that are used in differing frequencies. Additionally, they also provide general insight into the context in which these metrics are used.

Table 3.1: Overview of metrics compared and represented in van Vliet et. al (2016) and Tong and Feng (2020).

Metric	Occurences in Tong and Feng (2020) [percentage]	Occurences in van Vliet et. al (2016) [percentage]
Overall Accuracy	113 [32.6%]	1 [0.9%]
Kappa	107 [30.8%]	13 [11.4%]
Number of Patches	41 [11.8%]	-
Allocation Disagreement	38 [11.0%]	2 [1.8%]
Quantity Disagreement	38 [11.0%]	2 [1.8%]
Relative Operating Characteristic	33 [9.5%]	17 [14.9%]
Figure of Merit	29 [8.4%]	8 [7.0%]
Percentage Landscape	24 [6.9%]	-
Landscape Shape Index	23 [6.6%]	-
Mean Patch Area	20 [5.8%]	1 [0.9%]
Fractal Dimension	20 [5.8%]	1 [0.9%]
Kappa Location	20 [5.8%]	1 [0.9%]
Moran's I	19 [5.5%]	1 [0.9%]
Kappa Simulation	-	5 [4.4%]
Fuzzy Kappa	-	3 [2.6%]
Total Edge	-	1 [0.9%]
Clumpiness	-	1 [0.9%]
Kappa Histo	-	1 [0.9%]

3.2.4 Metric Typologies

It is important to select similarity metrics based on the reason that the comparison is made (Stehman, 1997). For the purpose of using land-use change models in a scenario discovery context, this requires better understanding of the effect that similarity metrics have on the clusters of maps that they lead to. While this research creates a first step in improving the understanding in this regard, it is only a start. However, other attributes of similarity metrics can also be identified that are likely to be important factors when deciding on which similarity metrics to use. Table 3.2 provides an overview of three attributes that have been identified based on experience gained through the literature search and processing land-use maps for clustering.

The calculation method is of interest as landscape metrics are much quicker to compute. The value of this increases as the number of maps that are compared rises. However, this does limit the set of metrics that are available. Commonly used metrics such as kappa and the components of difference defined by Pontius are all cell-based.

Whether a categorical metric is selected is based on the purpose of the analysis. Categorical metrics are likely more useful for highlighting specific land-use categories, even in the scenario discovery process. Nonetheless, there are numerous categorical metrics that can also be aggregated to contain multiple, or all land-use categories. Which of these options should be chosen is up to the user. For simplicity's sake

Table 3.2: Overview of identified metric attributes with regards to scenario discovery application

Attribute	Attribute Levels	Description
Calculation Method	Cell - Landscape	Pair-wise comparison (cell-based) or individual map metrics (landscape) impact the kind of information garnered and the computation time
Categorical Scale	Yes - No	Metrics can be calculated on a global scale, for a specific land-use category, or aggregate categories
Interpretability	Relative - Simple	All metrics have a given purpose behind their equation, yet not all outcomes are equally interpretable

categorical metrics are labeled as a binary choice, which is also more in line with how the metrics are generally used.

The final point of interest pertains to the interpretability of metric outcomes. This goes hand in hand with trying to understand what the effect is of similarity metrics on the formation of map clusters and the results that are obtained from applying a chosen rule induction algorithm afterwards. The chosen attribute levels are 'simple' and 'relative'. In this context relative means that the metric is likely hard to interpret without comparing it with another outcome so that relative values can be assigned.

Let us compare the metrics overall accuracy and kappa. Overall accuracy is a simple and intuitive metric that highlights the agreement between maps. On the other hand, kappa builds on what overall accuracy does and corrects for the inherent chance of randomly allocating cell values. While this could be a useful addition for validation purposes, it is unclear what the value means, except relative to another kappa outcome. For example, an overall accuracy value of 0,7 means that 70% of the cells between the compared maps agree on allocation of land-use. A kappa value of 0,7 would generally be interpreted as 'good' with regards to agreement, and better than the same value of overall accuracy if we are concerned about the model randomly assigning values (debatable whether this is of interest) but that is as far as it goes. Interpretability of metric values is then of interest as they are the basis for forming clusters, and provide a first clue as to what kind of clusters might form.

3.3 Metric Selection and Discussion

Following the literature review, 10 metrics were selected to serve as a basis for clustering the maps generated by the Land Use Scanner. This includes two categorical metrics for which the calculations are repeated for a total of three categories. The set of metrics are presented in Table 3.3.

One of the selection criteria for the similarity metrics is a diversity of cell-based and landscape based metrics. Since the purpose of the Land Use Scanner model is to observe global patterns of land-use change most metrics were chosen to observe such behaviour. As some land-use change models are created with more specific patterns in mind, categorical metrics were also added. Categorical metrics are those metrics that calculate their statistic for only a single land-use category when there are multiple land-use categories to choose from. This is based on the assumption that categorical metrics will be better able to capture changes when specific land-use categories are of interest.

Finally, interpretability of the metric was also considered for selection. In general a preference was given to those metrics where it is easier to interpret the value. The minor outliers here are kappa, Shannon's index, and Simpson's index. When clustering on a similarity metric it can quickly become challenging to understand what these clusters truly consist of. As this research is the first to explore the effect of similarity metrics, the choice was made to focus on similarity metrics that are easier to interpret the value of.

In the following sections each of the metrics will be presented by discussing how they work, what this might mean when applied to the data set, and whether any adjustments are made to the metric before use in clustering. An overview of all the symbols used to calculate the metrics, and a reference to the relevant equations can be found in Table 3.4. The first metric presented is kappa, which includes an example calculation to provide insight in how the process from a set of maps to clusters of maps works.

Table 3.3: Overview of selected similarity metrics

Metric	Description	Calculation Method	Categorical Scale	Interpretability	Application Examples

Kappa	Provides ratio of agreement between two maps corrected for random chance	Cell	No	Relative	Jin and Mountrakis (2013);
					Harvey et al. (2019);
					Liu, Zheng, and Wang (2020)

Overall Accuracy	Provides ratio of agreement between two maps	Cell	No	Simple	Jin and Mountrakis (2013);
					Kityuttachai, Tripathi, Tipdecho, and Shrestha (2013);
					Liu et al. (2020)

Total Difference	Sum of quantity and allocation difference	Cell	No	Simple	-

Total Quantity Difference	Difference across all categories in the number of each category assigned between two maps	Cell	No	Simple	Chaudhuri and Clarke (2014);
					Yalew et al. (2016);
					Aguejdad, Houet, and Hubert-Moy (2017)

Total Allocation Difference	Difference across all categories where the quantity allocated is the same but their location varies	Cell	No	Simple	Celio, Koellner, and Grêt-Regamey (2014);
					Gaudreau, Perez, and Drapeau (2016);
					Yalew et al. (2016);
					Aguejdad et al. (2017)

Percentage Landscape	Percentage of a specific category allocated relative to the total number of cells	Landscape	Yes	Simple	Ash et al. (2021); Cimatti et al. (2021)
Total Class Area	Total number of cells allocated to a specific category	Landscape	Yes	Simple	-
Quantity Difference	Difference for a specific category in the number of assigned cells between two maps	Landscape	Yes	Simple	See Total Quantity Difference
Shannon's Diversity Index	Represents the number of different elements in the data	Landscape	No	Relative	Punia, Joshi, and Siddaiah (2021); Cimatti et al. (2021)
Simpson's Diversity Index	Probability that two different elements belong to the same category	Landscape	No	Relative	Hu, Wang, Wen, and Xia (2012); Bibi, Ali, et al. (2013)

3.3.1 Kappa

Kappa is a commonly used metric that corrects the observed agreement by the probability of this happening by chance. The formula for Kappa can be found in equation 3.1, with its components calculated in equation 3.2 and equation 3.3. Kappa is a cell-based metric calculated from the contingency table. The main premise of kappa is that the observed agreement, as used in overall accuracy, is corrected by the chance of this allocation of observations being assigned by chance, given the frequencies in the observations.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3.1}$$

$$p_o = \sum_{j=1}^{J} C_{jj} \tag{3.2}$$

$$p_e = \sum_{i=1}^{J} \left(\frac{C_{ij} \cdot C_{ji}}{N} \right) \tag{3.3}$$

An example of a kappa calculation, and how this leads to clustering of maps, follows using the three example maps in Figure 3.1. Map 1 can be considered the base map where the changes on the other two maps are surrounded by black borders. Map 2 has two changes in the top right and map 3 has two changes in the bottom left. Calculating kappa for each of these pairs of maps leads to the values presented in Table 3.5.

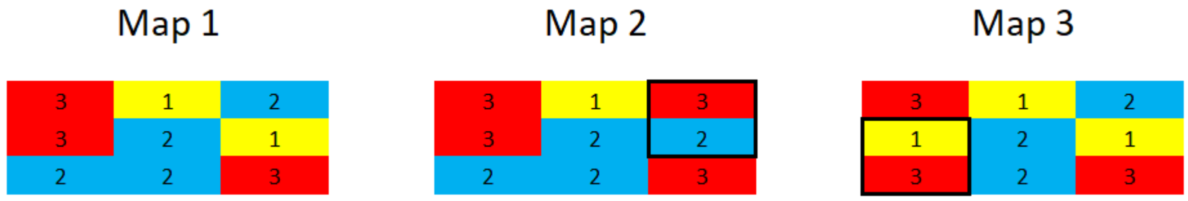


Figure 3.1: Base map with two different maps to use for comparison

According to these outcomes both map 2 and 3 are relatively similar whereas map 2 and 3 are very different. Visual inspection confirms this as both map 2 and three share agreement on seven counts with

Table 3.4: Overview of symbols and equations in which they are used

Symbol	Description	Calculated in eq	Component of eq
κ	Kappa metric calculated between two maps	3.1	
p_0	Observed agreement between two maps	3.2	3.1 3.4
p_e	Expected agreement between two maps	3.3	3.1
OA	Overall accuracy metric calculated between two maps	3.4	
Q	Total quantity difference across all categories	3.5	3.12
q_j	Quantity component for category j	3.6	3.5 3.11
A_p	Total allocation difference across all categories	1.7	1.12
S	Total shift component across all categories	1.8	1.7
E	Total exchange component across all categories	1.9	1.7
e_j	Exchange component for category j	1.10	1.9 1.11
s_j	Shift component for category j	1.11	1.8
D	Total difference across all categories	1.12	
d_j	Difference component for category j	1.13	1.11 1.12
PLAND	Percent of landscape occupied by a given category	1.14	
TCA	Total Class area	1.15	
H	Shannon's Diversity Index	1.16	
D	Simpson's Diversity Index	1.17	
N	Total number of cells		1.4 1.17
C_{ij}	Total number of cells where map A predicts category i and map B predicts category j		1.3 1.6 1.13
C_{jj}	Total number of cells where both map A and B predict category j		1.2 1.10 1.13
p_j	Proportion of landscape occupied by category j		1.16
a_{ij}	Area of patch ij		1.14 1.15
A	Total landscape area		1.14
n_i	Total number of cells occupied by category i		1.17

Table 3.5: Kappa values for the maps from Figure 3.1

Observation 1	Observation 2	Kappa
Map 1	Map 2	0.65
Map 1	Map 3	0.67
Map 2	Map 3	0.33

map 1 whereas they only share five out of nine counts with each other. However, five out of nine is still considerably higher than the kappa value represents. The question that is thus raised is, when applied to larger maps, what is the value of correcting for random chance?

Additionally, using kappa provides no information on where or how the agreement or disagreement between data sets occurred. In the case of this example, a change from land-use class 1 to class 2 may not be as meaningful as a change from class 2 to class 3. This kind of information is lost when using kappa.

Before these outcomes are used for clustering, we have to consider whether it provides the desired information. Since kappa is a similarity metric, a higher value indicates increased similarity between two maps. For clustering, we are however interested in minimizing distance between potential clusters. As such, we cluster on 1 - kappa, leading to the distance matrix in Table 3.6. When this information is used as a basis for clustering map 1 would first cluster with map 3, followed by map 2. However, after clustering once, part of the table would change as map 1 and map 3 are no longer separate entries, but form a cluster. This updated distance matrix after one cluster step is illustrated in Table 3.7. In this case the new distance value between the cluster and the final map is computed as the average distance between all maps in the cluster and map 3. This shows that the dissimilarity of map 3 relative to the new cluster changes as it was more similar to map 1 than to map 2.

Table 3.6: Distance matrix of the Kappa example (1-Kappa)

	Map 1	Map 2	Map 3
Map 1	0	0,35	0,33
Map 2	0,35	0	0,67
Map 3	0,33	0,67	0

Table 3.7: Distance matrix after first cluster step and new distance being average cluster distance.

	Map 1_3	Map 2
Map 1_3	0	0,51
Map 2	0,51	0

3.3.2 Overall Accuracy

Overall accuracy is a simple metric that shows the ratio observed agreement and the total number of cells between two maps as shown in equation 3.4. It is a simpler form of kappa as it also uses p_e, but without correcting for cell selection by chance. Unlike kappa, its value is very intuitive to interpret, being the ratio of cells on which both maps agree without any additional components. To cluster based on overall accuracy the complement of the metric has to be taken to find the dissimilarity. As with kappa, this means the distance matrix is computed as 1 - overall accuracy.

$$OA = \frac{p_o}{N} \tag{3.4}$$

3.3.3 Total Quantity Difference

Total quantity difference is a component of difference as defined by Pontius and Santacruz (2014). The metric provides insight in the different quantities of land use classes assigned by map A and map B. Pontius and Santacruz express the value of the metric as a proportion, allowing for comparison across classes. In this research the absolute values are used instead as no cross-class comparisons are made. As such the relevant equations for total quantity difference (equation 3.5) and categorical quantity difference (equation 3.6) are defined accordingly in this research. Because quantity difference already calculates dissimilarity, clustering will be done on the metric value instead of a correction of it, as with kappa and overall accuracy.

It is likely that quantity difference will provide very different results compared to kappa and overall accuracy because it focuses specifically on a term of difference between two maps. Where kappa and overall accuracy summarize the similarity between two maps in one value, quantity difference expresses the difference in the number of cells of each category. In this research q_j is also used as a metric for clustering. Unlike total quantity difference it allows for focusing on a specific category of interest. The expectation is that clusters formed with this metric can capture the impact of policy levers likely to impact specific land-use categories.

$$Q = \frac{\sum\limits_{g=1}^{J} q_j}{2} \tag{3.5}$$

$$q_g = \left| \sum_{i=1}^{J} (C_{ij} - C_{ji}) \right| \tag{3.6}$$

3.3.4 Total Allocation Difference

Total allocation difference was first defined as allocation disagreement (Pontius Jr & Millones, 2011) and later split into the components of shift difference (equation 3.8) and exchange difference (equation 3.9) in a later article (Pontius & Santacruz, 2014). Where quantity difference looks at a difference in number of cells assigned, allocation difference focuses on where each cell has a given class. As such, it is possible for a map to have no quantity difference, for example both map A as well as map B have 100 cells of land-use class 1, yet there can still be allocation difference because map A and map B might not have assigned each instance of class 1 to equivalent cells.

In this research total allocation is calculated as the sum of its components shift and exchange, yet these components are not used as a basis for clustering. Similar to quantity difference, the equations are not normalized to compute the proportion of allocation difference. The value of allocation difference is used directly for clustering as it can be considered a measure of dissimilarity.

$$A_p = E + S \tag{3.7}$$

$$S = \frac{\sum\limits_{j=1}^{J} s_j}{2} \tag{3.8}$$

$$E = \frac{\sum\limits_{j=1}^{J} e_j}{2} \tag{3.9}$$

$$e_j = 2 \cdot \left\{ \left[\sum_{i=1}^{J} MINIMUM(C_{ij}, C_{ji}) \right] - C_{jj} \right\} \tag{3.10}$$

$$s_j = d_j - q_j - e_j \tag{3.11}$$

3.3.5 Total Difference

Total difference combines the components of difference calculated by allocation and quantity into one metric (Pontius & Santacruz, 2014). This is shown in equation 3.12. It can also be calculated as the complement of the proportion correct (see equation 3.13).

Where quantity disagreement and allocation disagreement focus on one specific type of disagreement, total difference combines these two aspects. While it provides a better picture for total disagreement in this regard, as an aggregate metric it provides less information on what kind of disagreement occurs. Total disagreement between two comparisons might share the same value for total difference, while each results from different components of total difference. This could lead to problems in later stages of the analysis when maps are clustered based on similarity metric comparisons, when in reality this is undesirable.

$$D = \frac{\sum_{j=1}^{J} d_j}{2} = Q + A_p \tag{3.12}$$

$$d_j = \left[\sum_{i=1}^{J} (C_{ij} + C_{ji})\right] - 2 \cdot C_{jj} \tag{3.13}$$

3.3.6 Percentage of Landscape

Percentage landscape is another straightforward metric, and the first metric that is fully categorical in nature. It is calculated by taking all patches or cells of a specific land-use category and dividing that by the total area, as shown in equation 3.14. Another difference is that it is the first of the landscape metrics that is discussed, and thus has a unique metric value for a given map, instead of having a value for a pair of maps. The unique values of map 1 must thus be directly compared with map 2 and map 3, instead of having to compare map pairs.

A categorical metric such as this might be of interest when specific land-use classes are the focal point of study. One of the potential pitfalls of earlier metrics was not knowing where the terms of disagreement come from. Even though percentage of landscape does not provide information on location errors, it does allow for narrowing it down to specific categories of interest.

$$PLAND = P_i = \frac{\sum_{j=1}^{n} a_{ij}}{A} \cdot 100 \tag{3.14}$$

3.3.7 Total Class Area

Total class area is very similar to percentage of landscape, except that absolute area of the class is used instead of the class area relative to the total area, as can be seen in equation 3.15.

What makes total class area interesting, in addition to percentage of landscape, is because it provides a similar metric that might lead to different clusters. Both Percentage landscape and total class area calculate unique statistics for each map, instead of for map pairs such as the cell-based metrics. Because of this the distance value used for clustering is not inherent in the metric itself, but has to be computed afterwards. Where percentage landscape provides values relative to the total area, total class area does no such thing.

$$TCA = \sum_{j=1}^{n} a_{ij} \tag{3.15}$$

3.3.8 Shannon's Diversity Index

Shannon's diversity index (Shannon, 1948) is a measure often used in ecological literature to measure species diversity. The way it is calculated is shown in 3.16. The logarithm base can be changed, however should be consistent between comparisons. Its purpose is to make the result easier to interpret.

While not commonly used for general land-use change modelling, this metric was selected because its concept is still relevant. Instead of species diversity, the diversity of land-use allocation is measured. Since the Land Use Scanner model that is used focuses on a variety of land-use categories, the application of a general diversity metric is of special interest as it might be able to capture different aspects of the data than traditional similarity metrics.

$$H = -\sum_{i=1}^{s} p_i \cdot \ln(p_i) \tag{3.16}$$

3.3.9 Simpson's Diversity Index

Simpson's diversity index (Simpson, 1949) is another measure used in ecological literature to quantify species diversity. However, unlike Shannon's index it is often considered more of a measure of dominance than diversity. This can be traced back to its method of calculation, as found in equation 3.17, where less represented categories will become less apparent if there are dominant categories.

Simpson's index thus makes an interesting addition to the set of similarity metrics as one could argue that measuring dominance is the opposite of measuring diversity. That is to say, as diversity increases, dominance reduces, and vice versa.

$$D = \frac{\sum_{i=1}^{s} n_i(n_i - 1)}{N(N-1)} \tag{3.17}$$

3.4 Optimal Cluster Numbers

After all maps are compared using similarity metrics, they are clustered using the similarity metric outcome as distance values. This was done in Python using the agglomerative clustering algorithm. The number of clusters for each metric were found using a combination of the Elbow Method and the Silhouette Score. The Elbow method recommends the number of clusters to use by plotting the explained variance against the number of clusters for a given data-set. The cut-off point is where the diminishing returns of increased variance explanation do not outweigh using additional clusters is the 'elbow' and is considered the recommended number of clusters in this method. The Silhouette Score is a measure of how well each member of a cluster belongs to that cluster and other clusters. Both methods were used so that they could complement one another.

As it is likely that the number of clusters selected impacts the results in later steps of the scenario discovery process it is important to carefully consider the number of clusters to select for each metric. The Elbow Method is used as a visual tool to find the recommended number of clusters. This is then compared to the Silhouette Score, which provides both numerical and visual representation of the results. In general, the Silhouette Score seemed to prefer smaller number of clusters at the cost of in-cluster similarity. For all cases the recommendation of the Elbow Method was followed as deviations based on the Silhouette Score would lead to relatively large increases in in-cluster variance.

The selected number of clusters and their respective Silhouette Scores are presented in Table 3.8. With values approaching 1 suggesting correct cluster assignment and -1 implying incorrect assignment, a number of metrics score well. The exceptions being kappa, overall accuracy, total difference, and total allocation difference all scoring below 0.30. For the first three a large score increase can be found by reducing the number of clusters to two, yet the Elbow Method showed that it is still preferable to have a larger number of clusters. This shows the value and importance of not blindly following a single method to find the number of clusters. An example for finding the cluster number of kappa is described in Appendix C.

Table 3.8: Overview of similarity metric attributes related to clustering and representative map comparison

Metric	Silhouette Score	Cluster Count
Kappa	0.26	7
Overall Accuracy	0.27	7
Total Difference	0.27	7
Total Quantity Difference	0.50	4
Total Allocation Difference	0.28	6
Percentage of Landscape [Residential]	0.96	3
Percentage of Landscape [Corn]	0.55	5
Percentage of Landscape [Nature]	1.00	3
Total Class Area [Residential]	0.96	3
Total Class Area [Corn]	0.55	5
Total Class Area [Nature]	1.00	3
Quantity Difference [Nature]	1.00	3
Shannon's Diversity Index	0.54	4
Simpson's Diversity Index	0.55	5

3.5 Analysis Framework

The analysis of the cluster data previously described is done in three ways. Figure 3.2 globally visualizes the steps already described (up to step 3), and those that will be discussed further now. A more detailed description on generating and processing Land Use Scanner output can be found in Appendix B. Steps 4 and beyond will be now discussed.

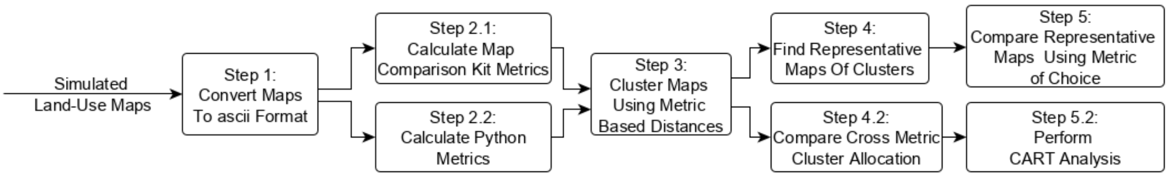

Simulated Land-Use Maps
Step 1: Convert Maps To ascii Format
Step 2.1: Calculate Map Comparison Kit Metrics
Step 2.2: Calculate Python Metrics
Step 3: Cluster Maps Using Metric Based Distances
Step 4: Find Representative Maps Of Clusters
Step 4.2: Compare Cross Metric Cluster Allocation
Step 5: Compare Representative Maps Using Metric of Choice
Step 5.2: Perform CART Analysis

Figure 3.2: Flowchart of data processing steps performed in Chapter 5

3.5.1 Representative Maps

The purpose of step 4 and 5 in Figure 3.2 is to find whether the resulting clusters, for each metric, contain different maps. To explore this, representative maps for each cluster were found. The representative map for each cluster is defined as that map of which the sum of its difference with all the maps of its cluster is the lowest. Although this is anything but a foolproof way of judging the distinctiveness of clusters, it does provide initial insight into whether the clusters found by using similarity metrics lead to different narratives based on land-use allocation.

The comparison of representative maps in step 5 is made using the normalized quantity difference metric (Pontius Jr & Millones, 2011). Note that any metric can be chosen, the metric of choice varying with the goal of the comparison. The reason that quantity difference was selected is that it is simple to calculate and interpret, while providing useful information. This metric, when normalized, provides the percentage of the total study area that map one and map two of the comparison assign differently from one another. This metric thus highlights a focus with regards to the quantity of each land-use category. What it does not provide is any information on whether the allocation of these categories is in the same location. For this comparison the absolute quantity was considered more relevant than either the allocation or the sum of both as a measure of disagreement.

To compare all representative maps for a given metric using the quantity difference metric, the first step is to store the quantity difference values for each representative map pair, before merging them into a single data set. A heatmap is then applied to the data set to give a visual indicator of where the most change occurs. The result is a data frame that contains percentile change of land-use allocation between maps, normalized to the total land area. This process is then repeated for all metrics in Table 3.9. This analysis helps create simple narratives for the representative maps, if they exist, and to thus judge if the different clusters contain distinct maps. Representative maps will be covered in Chapter 4. Finding the representatives maps, and the application and description for comparing the representative maps can be found in Appendix D.

3.5.2 Cluster Allocation

The cluster allocation analysis consists of confusion matrices for all metric combinations. These confusion matrices visualize the allocation of maps to a given cluster when using different metrics. Assuming an equal number of clusters between metrics, total agreement in cluster allocation would show maps spread across only as many labels as there are clusters. The higher this spread relative to the number of clusters, the bigger the difference between metrics in cluster allocation. This analysis is presented as part of Chapter 5. The full process of cluster allocation is described in Appendix E.

3.5.3 CART analysis

Classification and Regression Tree (CART) analysis is a decision tree learning technique that forms decision trees based on variable levels of the data. In this research these variables are the input variables used in the Land Use Scanner to generate different scenarios for the experimental setup. The tree comparison will consist of a visual analysis in combination with a comparison of tree attributes (Bošnjak, Karakatič, & Podgorelec, 2015). These tree attributes are:

- Maximum number of decisions taken
- Total decision nodes in the tree
- Number of different attribute used, including percentage showing relative to total number of attributes
- Number of different attributes between trees, including percentage showing relative to total number of attributes used for that metric

These four attributes add a structural component to the visual comparison of the CART trees. Other means of comparing decision trees were found in a literature search, yet no agreement was found on how to approach it. As such, this simple approach using attributes suffices for the scope of this research. CART analysis is part of Chapter 5. Details on the CART tree creation, and references to all created trees can be found in Appendix F.

3.6 Experimental Setup

This research uses 2,000 maps generated by the Land Use Scanner model as input for further processing. The generation of these maps is described in Cox (2020). How these results were processed into clusters varies based on the metric. Table 3.9 presents the relevant metric attributes.

Comparing the maps using each similarity metric was done with either Python or the Map Comparison Kit. The custom Python implementation calculates each similarity metric value and stores it as required. The Map Comparison Kit is a stand-alone program generally used to compare small numbers of maps, developed by the Research Institute for Knowledge Systems (RIKS). Using the Map Comparison Kit was also performed through Python, but requires a few more steps to work. Both tools and their implementation are discussed in more depth in Appendix A, including references to the code used.

A notable difference between similarity metrics is whether they are calculated as a unique value per map or require map pairs for computation. In the former case, each map requires only a single calculation after which basic arithmetic calculations can be used to compute the contingency matrix used for clustering. This is different for similarity metrics that require map pairs to compute values as each possible map permutation will have to be run. As the computational demand varies between similarity metrics this can end up being computationally demanding. Based on this attribute, a preference is found for map value similarity metrics over map pair metrics when possible.

When clustering using outcomes from similarity metrics (Step 3 from Figure 3.2), it is important to consider whether the similarity metric value is also a sensible distance value for the clustering algorithm. A clustering algorithm generally needs a distance value between observations, and in the case of hierarchical agglomerative clustering, it clusters bottom up starting with the lowest distance. When looking at the kappa metric, which calculates the similarity of two maps, clustering on this value the most dissimilar maps would be clustered first. As this is undesirable the complement 1 - kappa is instead used as the input for the clustering algorithm. This can be found in the clustering application column of Table 3.9.

Three of the metrics used are categorical in nature. These metrics are percentage landscape, total class area and quantity difference. A metric is considered categorical if its calculation is based on only one of the land-use categories available in the model. Since the Land Use Scanner model has over twenty land-use categories, only a sample of these were chosen for testing in combination with the categorical metrics. Since some land-use categories in the model are static, while others show varying degrees of outcome variance, three land-use categories were selected, where each category represents a category with either little, medium or high outcome variance. Outcome variance in this case refers to the number of maps in which a given land-use category has different land-use coverage. Low variance indicates that a lot of maps share a small number of values for a land-use category, whereas high variance means that most, if not all, maps have a unique outcome for the area covered.

An overview of these three metrics and their outcome variance can be found in Table 3.10. With a total of 2,000 maps used in this research, the final column in the table provides the number of different land-use allocation totals that were found for that land-use category across the 2,000 map set. For computational reasons, quantity difference was only calculated for the land-use category nature.

Table 3.9: Overview of similarity metric attributes

Metric	Calculation Tool	Computation Type	Clustering Application	Categorical
Kappa	Python	Map Pair	1 - Kappa	No
Overall Accuracy	Python	Map Pair	1 - Overall Accuracy	No
Total Difference	Python	Map Pair	As Is	No
Total Quantity Difference	Python	Map Pair	As Is	No
Total Allocation Difference	Python	Map Pair	As Is	No
Percentage of Landscape	Python	Map Value	As Is	Yes
Total Class Area	Python	Map Value	As Is	Yes
Quantity Difference	Python	Map Pair	As Is	Yes
Shannon's Diversity Index	Map Comparison Kit	Map Value	As Is	No
Simpson's Diversity Index	Map Comparison Kit	Map Value	As Is	No

Table 3.10: Overview of selected land-use categories and their outcome variance

Land-use category	**Outcome Variance**	**Number of varying maps**
residential	low	15
nature	medium	737
corn	high	1686

Chapter 4

Individual Metric Analysis

In this chapter the cluster results are analyzed separately for each metric using representative maps. First an overview is provided for how this was approached. Following this, the resulting map clusters are discussed for each similarity metric. The final section provides a conclusion to what was presented in the chapter.

4.1 Metric Cluster Analysis

In the following sections each metric will have its representative maps compared using the quantity difference metric. This provides initial insight as to whether different clusters contain sets of maps that might lead to different scenario narratives. Note that the representative map is only the map that is considered the most similar to all other maps in that cluster. The results in this section are thus a measure of what the other maps in the cluster are like, yet it does not provide conclusive information about the whole cluster.

The figures containing the quantity difference data will vary between the metrics. This is in part due to a varying number of clusters, leading to a different number of maps and thus rows. Additionally, insignificant values were removed. In this case that is considered a quantity difference of less than 0,0001% of the total area. This value is selected to keep a balance between interpretability, by not having too decimals, without neglecting the fact that these percentages represent a sum large enough that even fractions of a percentage constitute to a significant change in land-use allocation. Because of this cut-off some figures may have more columns than others as certain land-use classes may prove insignificant for the representative maps of a given metric.

All figures in the following sections only show the possible combinations for the representative maps. Although this provides the same information as the maps showing the permutations, it is far more compact at the cost of interpretability. Additional figures were also made to isolate each different land-use category. This provides useful information at a glance, but it was found that it makes it more difficult to characterize the nature of each cluster. As such, the presented figures, which are more difficult to analyze at a glance, are still used. References to the figures with all permutations, and the figures for each land-use category can be found in Appendix D.

4.1.1 Kappa and Total Difference

The first metric is the kappa metric. It will be covered together with total difference as both metrics provide the exact same representative maps, meaning the follow up analysis for both metrics is also the same. These metrics have seven clusters and thus seven representative maps to be compared. Eleven land-use classes show significant change between these representative maps. Figure 4.1 illustrates that the most significant differences in land-use class assignment are found in the residential and pasturage classes. Map 359 and map 1911 have a relatively higher assignment of residential area than the other maps. Industrial assignment is also higher, yet map 456 and map 695 also show higher assignment of this category. Map 359 also shows a unique increase in recreational and residential area, while map 456 is also the only map to show an increase in nature relative to the other maps. For all these maps, the increase in these categories comes largely at the cost of pastures. Based on these variations it can be said that the representative maps for kappa and total difference seem to highlight maps with a different land-use assignment focus.

There are two pairs of maps which are relatively similar, namely map 359 and 1911, and map 541 and map 1148. Whether or not this is significant depends to what percentage change can be considered relevant. For the maps 359 and 1911 the biggest difference is only 0.18% with regards to the construction area class. Relatively speaking this is much less than some of the other categories that are shown, yet it still means that map 359 predicts that 0,18% more of total Dutch land is dedicated to construction than map 1911. Further sections will continue to highlight this aspect where relevant, but, unless stated otherwise, it will be assumed that these differences are relevant.

Both kappa and total difference capture disagreement between two maps without focusing on specific aspects. As such it is hard to predict what might come out when they are used as a basis for clustering. It is of interest that only one of the representative maps of these metrics shows a variation with regards to nature allocation. Numerous levers on the Land Use Scanner impact the allocation of this category, yet it is hardly represented here. It is also of interest that one of the key categories in the model, namely residential land-use, is allocated across two different clusters even though they have the same value. This raises concerns for the metrics' ability to separate these results.

4.1.2 Overall Accuracy

Overall accuracy has seven different clusters with eleven significant land-use classes. Figure 4.2 highlights that for overall accuracy the focus is also on residential, industrial and pasturage categories. Between kappa and overall accuracy there is a difference in two of the seven maps, namely map 10 and map 696 are exchanged for map 1148 and map 1760. Because of this some similar trends return, such as map 359 and map 1911 showing high residential and industrial land-use allocation but a reduction of pasture land. For the new maps, both map 10 and map 696 show a reduction or close to equilibrium for both residential and industrial land-use classes, unlike most other maps. Together with map 359, map 10 is the only map to show a relative increase in recreational area in residential locations. Both map 10 and 696 seem to favor pasture land, however map 696 seems to assign it more consistently relative to other maps. Relative to one another, the biggest difference is that map 696 assigns more land to corn farming. Map 10 and 541, and map 359 and 1911 are the two pairs showing the least change in land-use allocation. Yet there are once again notable differences in land-use assignment between most of the representative maps.

Overall accuracy is a simple metric that shows, as a percentage, the cell-by-cell agreement. Since it does not focus on specific aspects of land-use, it is hard to predict what kind of clusters it might create. Similar to kappa the effect of the various nature related levers is lost with this metric. It is possible that this metric, which only focuses on global agreement, is useful only for the purpose of highlighting the large change that occurs in the model, instead of picking up on more subtle aspects. However, validating this statement would require a separate investigation into the variance between the representative map and the other maps in its cluster.

4.1.3 Total Allocation Difference

Total allocation difference has six clusters with nine significant land-use classes. Unlike the previous metrics, residential and industrial categories show little to no change for this metric. Instead, the focus

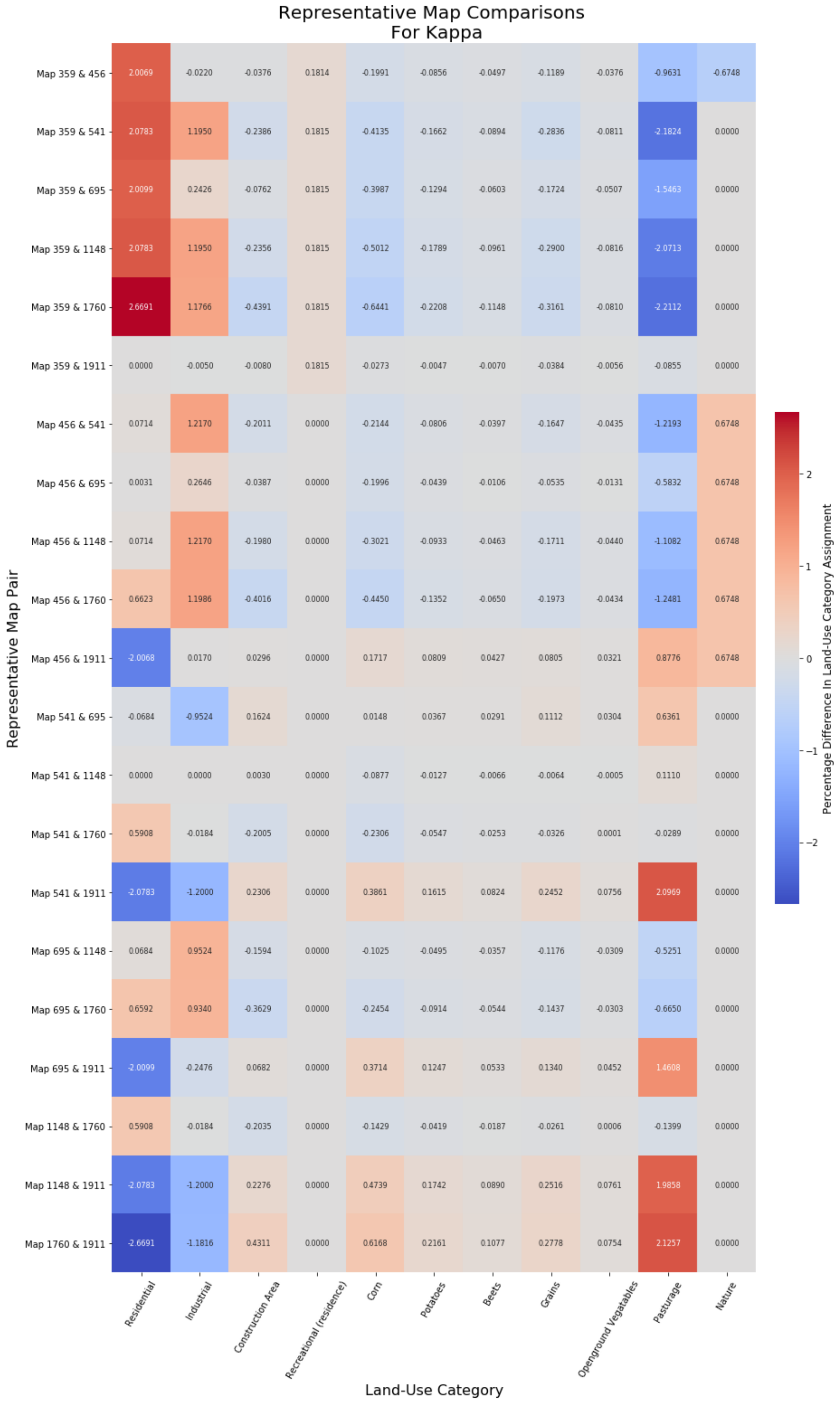


Figure 4.1: Quantity Difference for Kappa and Total Difference representative maps

is still on pasturage, while corn farming and construction areas show up more prominently. Map 1760 and map 1852 show clear decreases in pasturage, likely in favour of corn farming. Map 263, 453 and 1918 show the highest allocations of pasture land, yet it is not as high as the reduction for the other two maps. Map 263 is the only one that truly stands out when it comes to construction areas. For corn

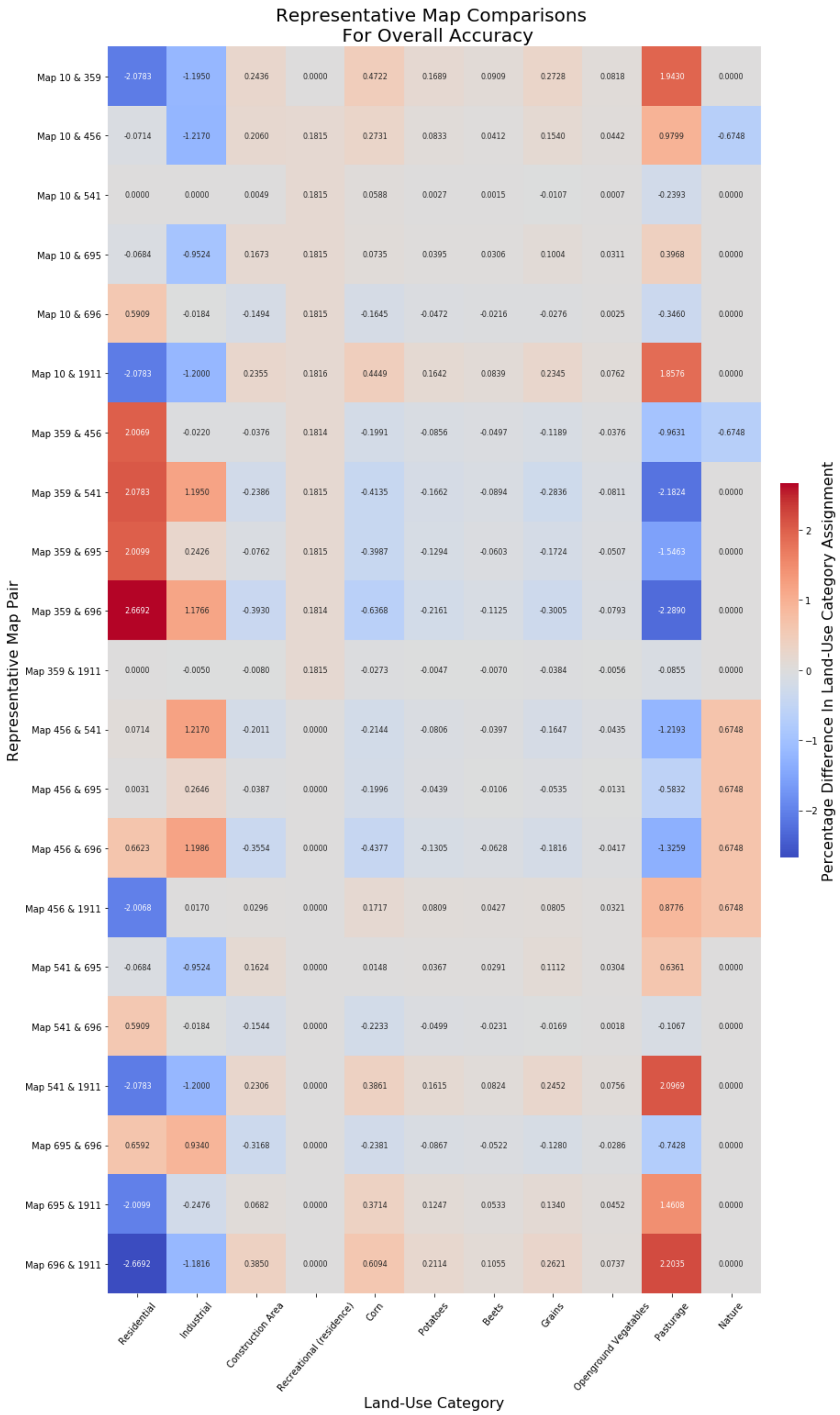


Figure 4.2: Quantity Difference for Overall Accuracy representative maps

there is quite some variation with map 263 and 1760 largely showing relative gains, whereas map 453, 678 and 1918 show the most relative decreases. Potatoes, beets, and grains see some small give and take between the maps whereas open ground vegetables largely show small differences. Nonetheless, the set of representative maps seems to show that different maps tend to different land-use assignment.

The results of total allocation difference are interesting as it is the first metric to put the focus on the nature land-use category. Since allocation difference is the switching of land-use from one location to another, this result is interesting as it might highlight how the various nature levers force nature land-use to move places instead of simply increase or decrease in quantity. The total lack of representative maps with high or low values for the land-use categories residential, industrial or pasturage means that the previous finding does come at the cost of other aspects. A metric such as allocation difference could prove to be useful as an additional metric to run to discover unexpected patterns.

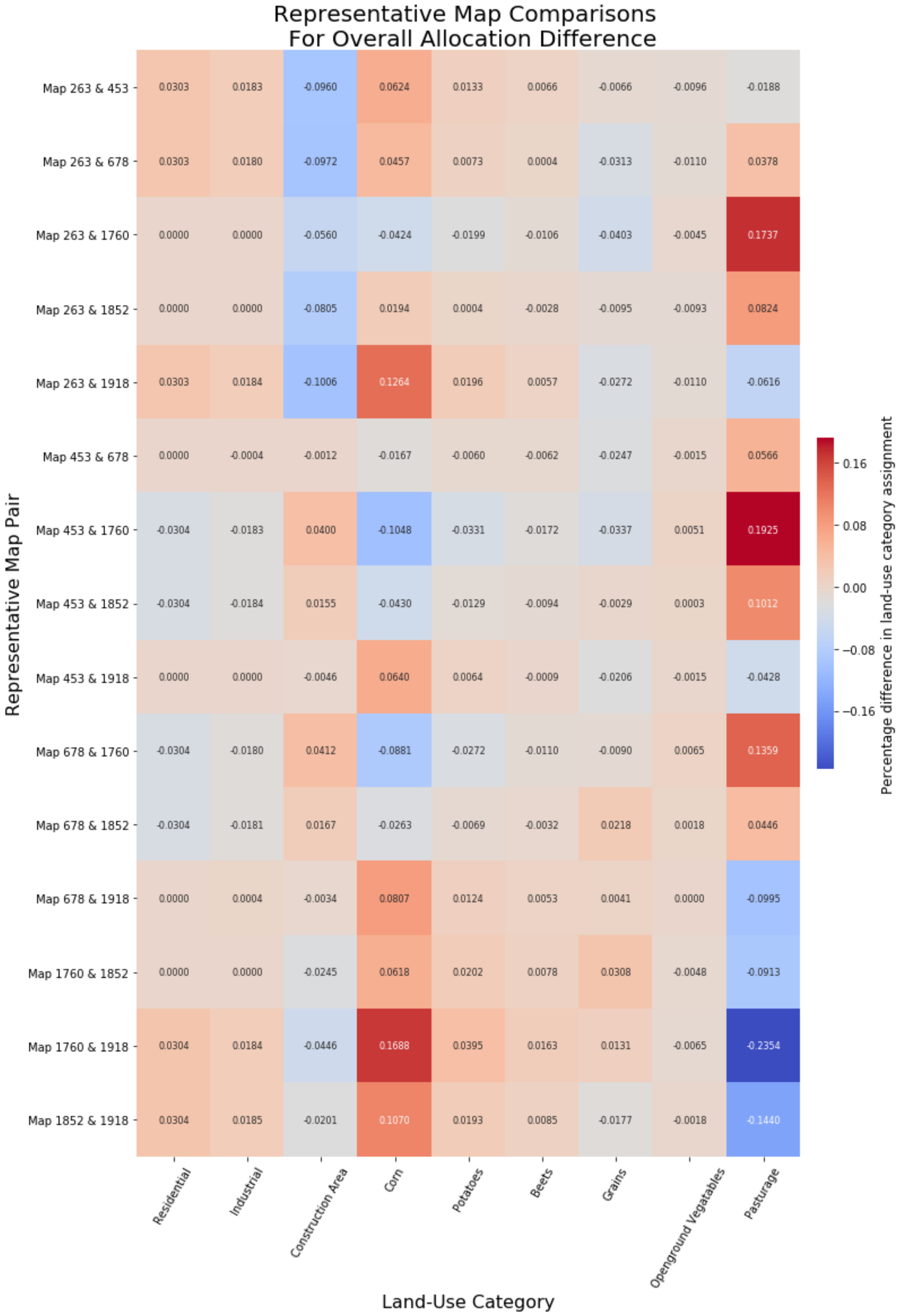


Figure 4.3: Quantity Difference for Total Allocation Difference representative maps

4.1.4 Total Quantity Difference

Total quantity difference has four clusters with fifteen significant land-use classes. Once again residential, industrial and pasturage categories show the most difference. For this metric nature also shows as a category with relatively high change. Map 1472 is a standout map that returns to the focus of increased residential and industrial assignment at the cost of mostly pasture land. Map 1833 sees the largest relative increase in nature land whereas map 1987 sees the largest decrease. Map 1978 seems to tend to more farming of all types at the cost of residential, industrial, and nature areas. For corn there are still minor patterns visible between the maps, which echo less distinctly into the potato, beets and grain land-use categories. For total quantity difference the representative maps also show different land-use assignments for each cluster.

Total quantity difference sums the quantity difference of different land-use categories together, because of this land-use categories with relatively higher change could eclipse less significant categories. The metric does seem able to better distinguish land-use categories between clusters. Residential, industrial, pasturage and nature categories seem to be split better across clusters than some of metrics have shown. This might be in part due to a lower number of clusters, but since the number of clusters is assigned based on suitability this is likely just a benefit of this metric.

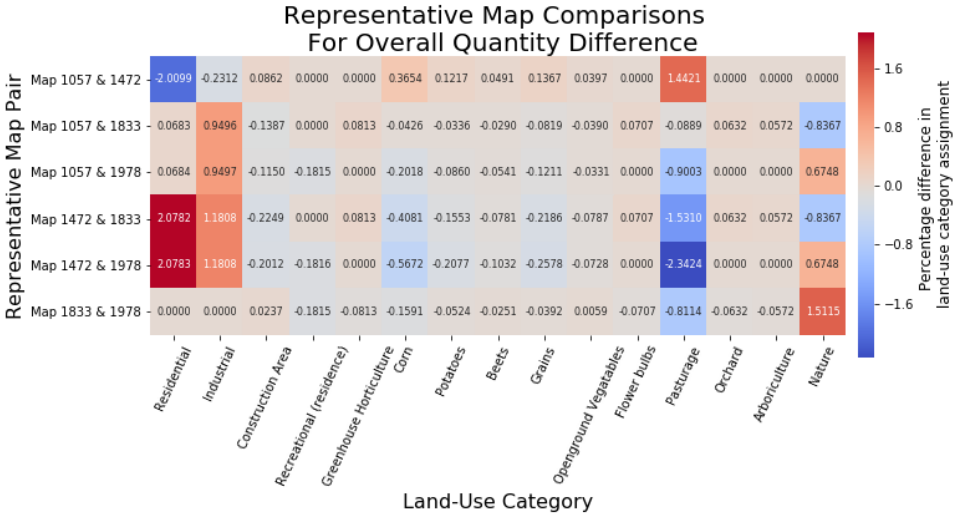


Figure 4.4: Quantity Difference for Total Quantity Difference representative maps

4.1.5 Shannon's Diversity Index

Shannon's diversity index has four clusters with fifteen significant land-use classes. Similar to previous metrics, residential, industrial, pasturage and nature land-use show the most distinct patterns, as shown in Figure 4.5. One difference between the two is that the extremes in the pasturage and residential categories are higher than any metric has shown thus far. In this set of maps, map 1457 is clearly the odd one out showing the largest difference with the other maps for most land-use classes. The other three maps are relatively similar, showing difference largely for either the pasturage or nature classes. This suggests that, unlike some of the previous metrics, Shannon's might have clustered many of the residential and industrial dominant maps together in one cluster.

Shannon's diversity index is commonly used as a measure of diversity. When applied for clustering it seems like this diversity is potentially translated into clustering all extremes together in a single cluster. This could be a useful attribute as further cluster analysis using this metric would then be on that specific extreme cluster.

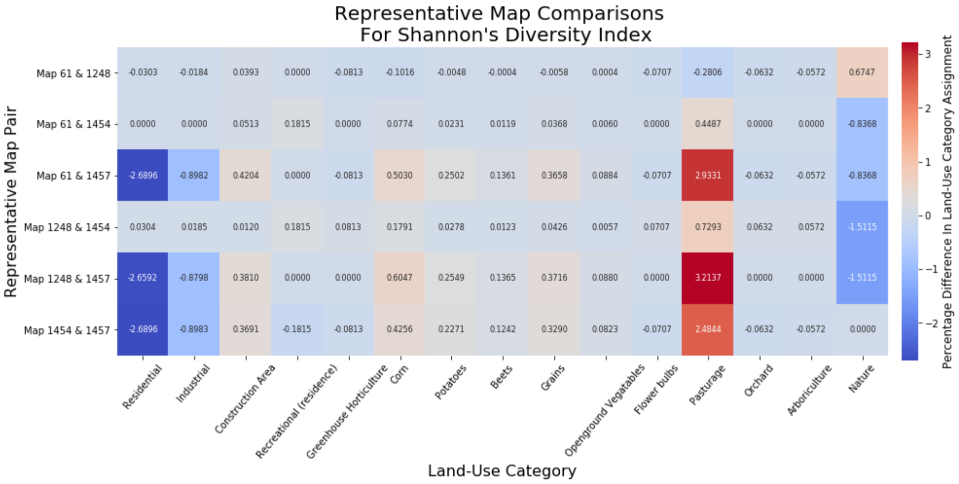


Figure 4.5: Quantity Difference for Shannon's Diversity Index representative maps

4.1.6 Simpson's Diversity Index

Simpson's diversity index has five clusters with fifteen significant land-use classes. Similar to Shannon's there is a focus on residential, industrial, pasturage and nature categories with once again some very noticeable extreme values, as displayed in Figure 4.6. Unlike Shannon's there is once again a return to variance among maps for the residential and industrial categories. Of interest are map 65 and 1830 which show less than 0,1% change across all categories apart from the construction and pasturage classes. Since it is unclear what the 'cut-off' margin between clusters is it is hard to say whether this is still a significant margin between the maps, even if it is largely small compared to other differences observed. Simpson's diversity index shows difference between representative maps as well, but two of the maps do appear relatively similar enough to constitute doubt.

Simpson's index is often considered a measure of dominance. On a categorical basis, the maps show a clear affinity for specific land-use categories. The exception to this is the nature category where there seems to be no clear pattern between the representative maps. The dominance aspect of this index returns in its focus on the land-use categories that see the most change in the model (residential and pasturage). While, large differences are also found in the nature category, the representation is likely an after effect of the more dominant categories. While an unusual metric for general land-use change modelling, this representation of dominance could be a useful attribute.

4.1.7 Total Class Area

Total class area is the first categorical metric to be presented. As it was analyzed for three different land-use categories, they will all be discussed sequentially. The three land-use categories are residential, corn, and nature.

Residential

Total class area for the residential category has three clusters with eleven significant categories. As visible in Figure 4.7, the focus is now on industrial, pasturage and nature land-use. With the extreme value being 1,19% we see the lowest difference for all metrics thus far. Map 217 has relatively little assignment of industrial land-use compared to the other two maps. Map 332 assigns less pasturage land whereas map 604 allocates the most. Additionally, map 217 and map 332 agree on nature allocation whereas map 604 allocates less than the other two. This first example of only three clusters being used shows a clear difference between representative maps.

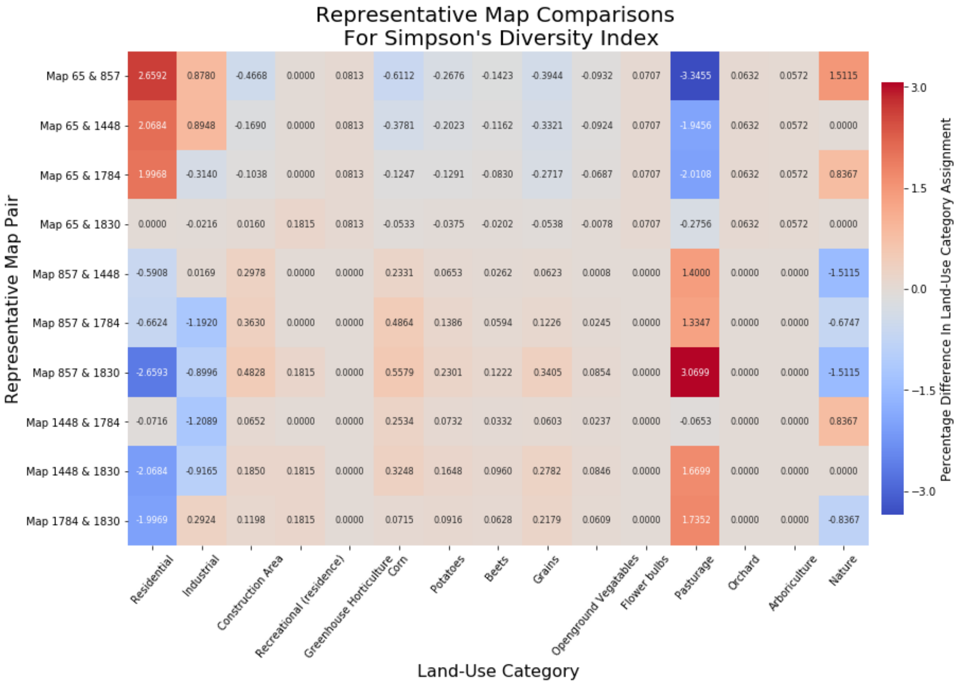


Figure 4.6: Quantity Difference for Simpson's Diversity Index representative maps

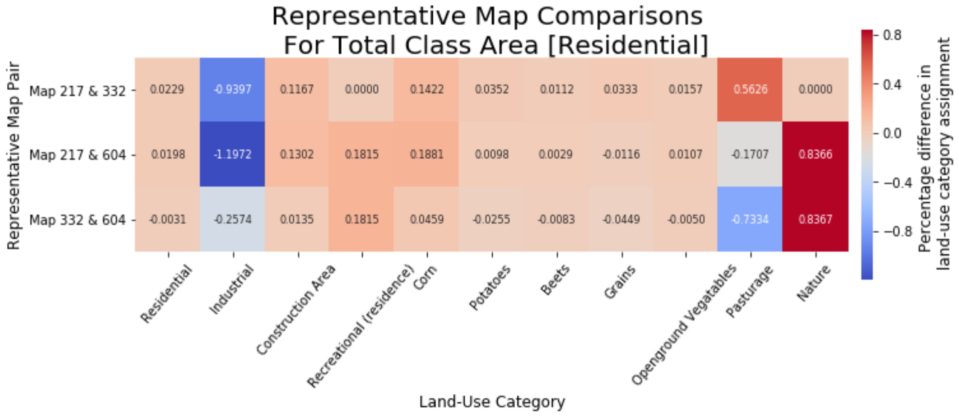


Figure 4.7: Quantity Difference for Total Class Area [Residential] representative maps

Corn

Total class area for the corn category assigns five clusters with fifteen significant categories (see Figure 4.8). Once again residential, industrial, pasturage and nature categories are most noticeable. Both map 246 and 325 have a strong residential focus, but map 325 has a higher allocation towards pasturage. Map 296, 825 and 1612 have similar residential allocation, and instead disagree on industrial, pasturage, and nature allocation. Map 296 and 825 both have relatively high corn allocation as well, which is repeated to a lesser degree for the categories potatoes, beets, and grains. There is considerable overlap between maps, but a difference in at least one category can always be found.

Nature

The final category for total class area is nature, which has three clusters with fifteen categories. Figure 4.9 illustrates how the representative maps for this category still show a difference for itself, the nature category, which was not the case for the residential category. Map 0 seems to represent a middle ground

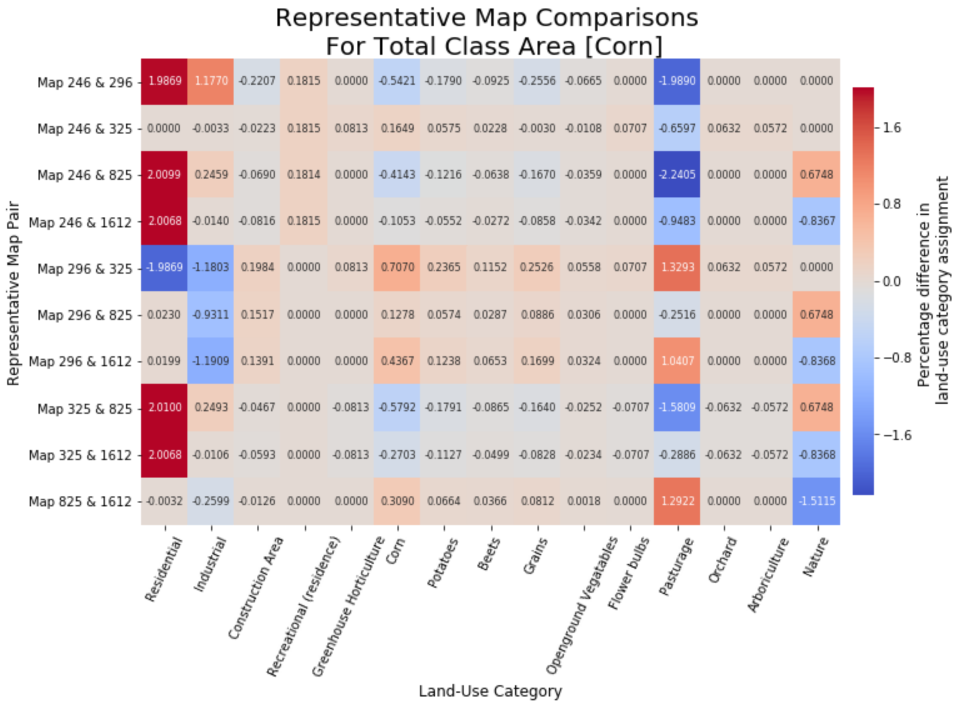


Figure 4.8: Quantity Difference for Total Class Area [Corn] representative maps

between map 8 and 35 when it comes to pasturage allocation, while also assigning less nature. Map 8 allocates less pasturage, but more nature than the other two maps. Map 35 represents less industrial and residential land for more pasturage than the other two maps, yet serves as a middle ground for nature allocation. Similar to the residential category, nature's representative maps also allow for a clear narrative difference between its representative maps.

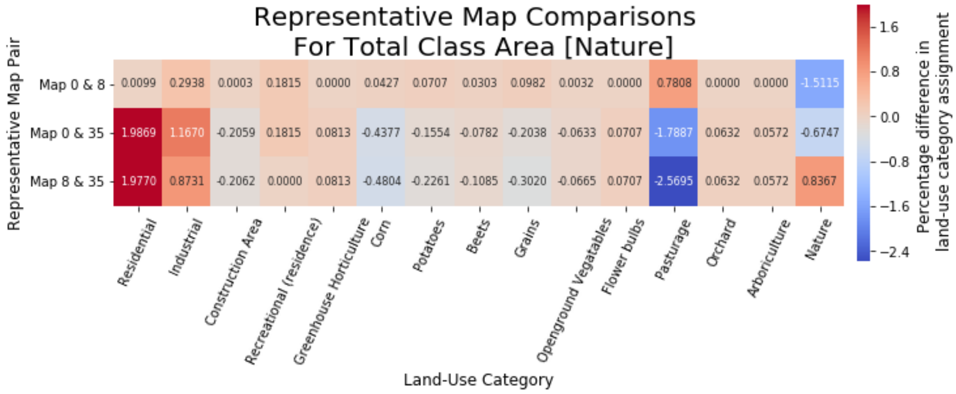


Figure 4.9: Quantity Difference for Total Class Area [Nature] and Percentage Landscape [Nature] representative maps

4.1.8 Percentage Landscape

Percentage landscape is the second categorical metric, being the total class area normalized by total land. It will be presented in a similar fashion as total class area, however, as the representative maps for the category nature overlap with that of total class area it will not be repeated. The section concludes with a comparison of the total class area and percentage landscape outcomes.

Residential

The residential category for percentage landscape has three clusters and eleven categories. The percentile differences between the representative maps are displayed in Figure 4.10. Map 264 shows less allocation than the other two maps for residential, industrial, and nature categories, yet an increase across all other categories, most notably pasturage and construction areas. Map 1422 and 1877 show little disagreement on residential land-use, but differences are seen for industrial, corn and nature land-use. With this it can be said that these representative maps are likely to be part of distinct clusters.

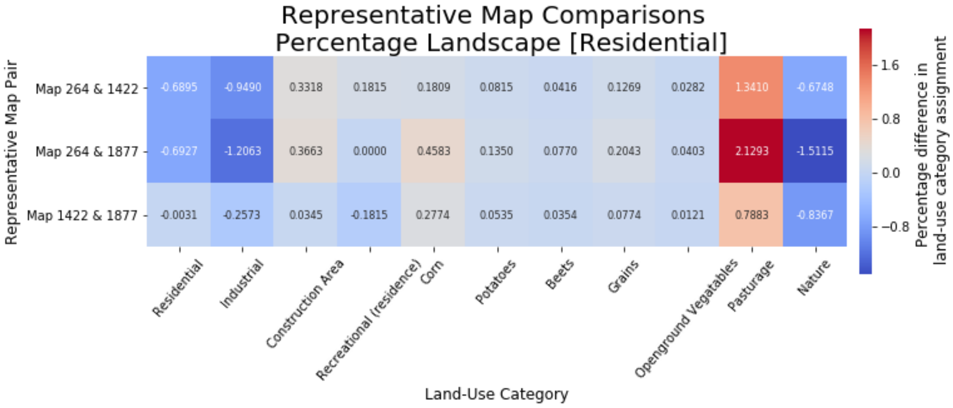


Figure 4.10: Quantity Difference for Percentage Landscape [Residential] representative maps

Corn

Percentage landscape for corn has five clusters and fifteen categories. Figure 4.11 shows the differences between the representative maps of these clusters. In this figure it is interesting that for the residential and industrial categories no map ever shows dominance, or subservience, unlike the pasturage and nature categories, where map 246, 704 and 843 take these roles. The exception to this is map 704 which allocates less industrial land-use than all other maps, albeit only barely less than map 843. Nonetheless, there are patterns visible such as map 704 and 843 both assigning little residential and industrial land, relatively high pasturage, yet assigning opposite with regards to nature. Similarly map 246 and map 325 assign more residential and industrial land but less pasturage. However, map 325 still assigns considerably more land to pasturage than map 246. Finally, map 564 sits somewhere in the middle on the big categories residential, industrial and pasturage, while assigning the same to nature as map 246 and 325. As such there is always some variation between maps across all categories.

Cross-class comparison for Total Class Area and Percentage Landscape

Comparing the different land-use category results it stands out that of all three categories, residential is the only category where the representative maps show little variation in the land-use category that they are clustered on. Similar to total class area, the residential category shows less variance in allocation than is observed for other metrics. It does show more variance for percentage landscape than it does for total class area. In general this is surprising, because intuitively clustering on a categories' total area would suggest that this would bring forth the different levels of land-use for the selected land-use category in the data set, and as such be found in the presented figures. As little to no change is observed for the selected category, this is clearly not the case. Either the residential category is the odd one out, or the clustering process proceeds differently than might be expected intuitively. It is also worth repeating that the other two categories, corn and nature do show up more prominently in their analysis, considering that most metrics were dominated by a combination of land-use change in the residential, industrial and pasturage categories.

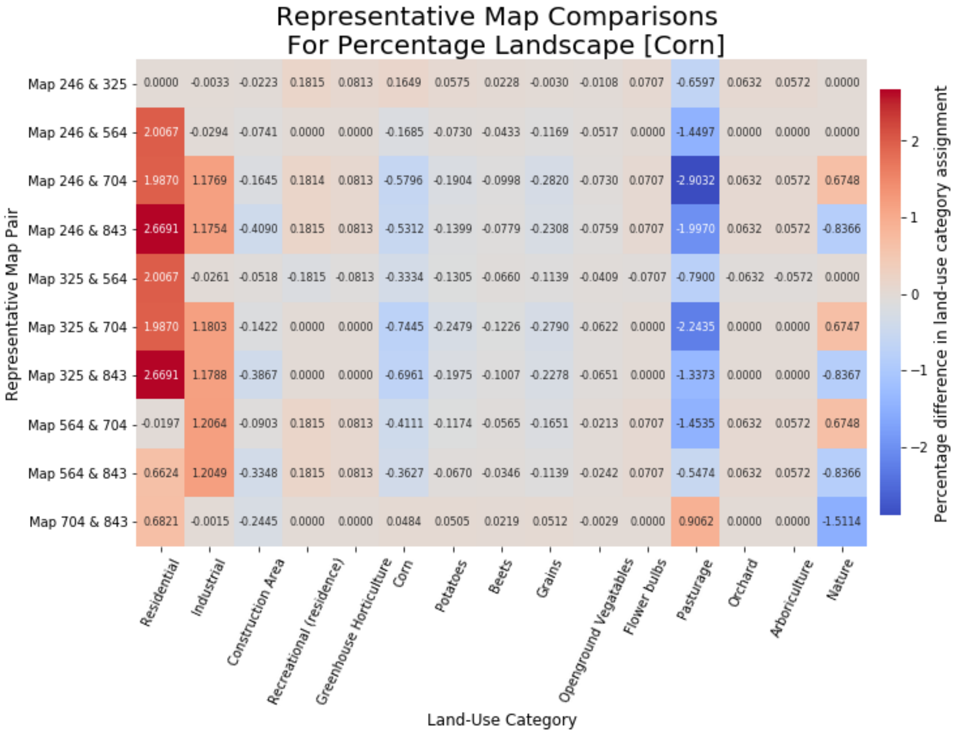


Figure 4.11: Quantity Difference for Percentage Landscape [Corn] representative maps

4.1.9 Quantity Difference

The final metric to be presented is quantity difference for the land-use category nature. As shown in Figure 4.12 it has three clusters with fifteen representative maps. Map 0 has a tendency to high residential and industrial land use at the cost of pasturage and nature. Map 1 is the opposite with relatively low residential and industrial land-use allocation but high pasturage and medium nature allocation. Map 8 is similar to map 0 but assigns even less pasturage and allocates more instead of less nature. Relative to map 1 the other two maps also allocate less land to farming other than pasturage. The representative maps for this metric also suggest that the different clusters contain different sets of maps.

Quantity difference for the nature category is interesting as the representative maps all stand out for different land-use categories. As the Land Use Scanner contains a variety of nature related uncertainties, it is unclear whether the variation in the other land-use categories is a coincidence or explicitly due to the metric itself. That is to say, each cluster as a whole might show a lot of variance for each land-use category, besides the nature allocation which is then similar within a given range. As the representative map is most similar to all other maps in its cluster it is also likely that the selected maps are more moderate than those that are in the cluster. As an example, map 0 shows the lowest nature allocation amongst all three maps, yet it is likely that its cluster contains maps with even less nature allocation.

4.2 Chapter Conclusion

This chapter compared the representative maps of each similarity metric on a map-by-map basis. The comparison of representative maps is made using the quantity difference metric (Pontius Jr & Millones, 2011). This metric computes the difference in total land assigned to each land-use category between two maps. It is calculated in such a fashion that it does not incorporate the difference in allocation of land-use to the same areas. As such, allocation is considered a different component of disagreement and there are differences and similarities that are thus missed. This is an inevitable result of using metrics. Nonetheless, the analysis showed that, as far as the representative maps go, the clusters for each metric

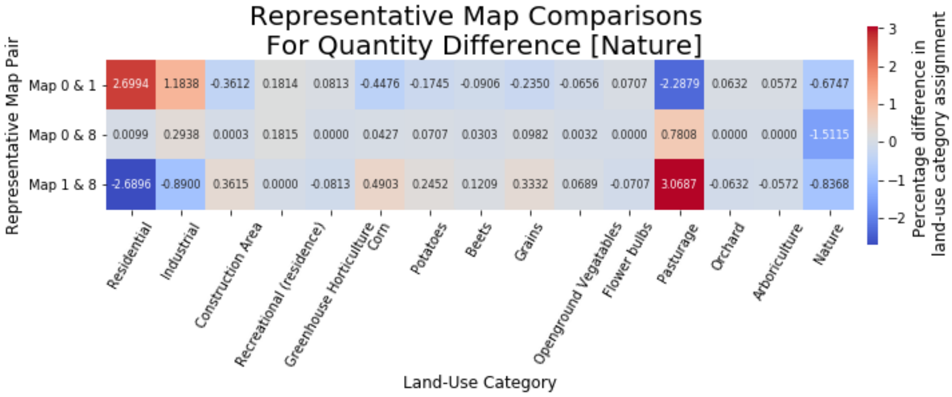


Figure 4.12: Quantity Difference for Quantity Difference [Nature] representative maps

seem to have their own distinct set of maps. Across metrics it was apparent that for the Land Use Scanner model the main land-use categories of change were residential, industrial, pasturage and nature categories. However, the different clusters often showed some overlap between one or more land-use categories yet variation in the remaining ones. Unsurprisingly, this is more likely when a metric was analysed with more clusters.

The analysis also showed some initial differences between the metrics. Table 4.1 highlights some notable attributes that the analysed metrics might have. While the performed analysis is far from sufficient to confirm such attributes, the presented results in combination with the mathematical background of each of these metrics do strengthen these suspicions.

Table 4.1: Summation of potential attributes of interest for specific metrics

Similarity Metric	Notable Attribute
Kappa Total Difference Overal Accuracy	Only capable of highlighting large changes with no care for how they occur
Total Allocation Difference	Capable of highlighting spatial relocation of land-use even when the relative share of land-use is low
Total Quantity Difference	Capable of identifying (large) differences in land-use allocation
Shannon's Diversity Index	Capable of creating a cluster of extreme results
Simpson's Diversity Index	Capable of highlighting dominant land-use categories
Total Class Area Percentage Landscape Quantity Difference	Categorical nature allows for highlighting change in specific categories Not fully reliable as shown by the residential land-use category

Chapter 5

Cross-Metric Analysis

This chapter continues where the previous chapter left off by comparing clustering results between similarity metrics. First the representative maps are presented. Afterwards a subset of comparisons will be discussed. Following this will be a global comparison of all CART trees using two different approaches. The chapter ends with a conclusion.

5.1 Representative Maps

The representative maps first presented in Chapter 4 to compare the distinctiveness of clusters for each metric are highlighted again to provide a first look into the cross-metric differences when clustering on similarity metrics. Table 5.1 presents all metrics with their representative map for each cluster. The results show a difference in number of clusters for most metrics, varying from three to seven clusters. A variance in cluster count is likely to result in different narratives, but this is currently unexplored.

It stands out that kappa and total difference share exactly the same maps. kappa and overall accuracy disagree on two out of seven maps. Both of these pairs will be discussed in depth in sections 5.2 and 5.3 respectively. For the non-categorical metrics there is no further similarity in representative maps, and generally in the number of clusters either. For the categorical metrics, which are those metrics that are calculated with the focus being one on land-use category as highlighted in brackets in the table, there is a complete overlap in representative maps for percentage landscape (PLAND) and total class area (TCA) for the nature land-use category, to minor overlap (corn) and complete disagreement on representative maps (residential). These differences between PLAND and TCA and the different land-use categories are also explored further in section 5.4

5.2 Kappa and Total Difference

Kappa and total difference are an interesting pair as they seem to lead to the exact same results. This was first hinted at in Chapter 4 where the representative maps were exactly the same. This suspicion is further confirmed by comparing how each metric allocates their maps to different clusters. Where Table 5.1 showed that both metrics use seven clusters, Figure 5.1 also shows that there is perfect agreement in allocation of maps to clusters between the two metrics. There is not a diagonal confusion matrix, however this is the result of how Python processes the clustering results. As there are only seven squares filled in the matrix with seven clusters for both maps, there is in fact perfect agreement.

This result is surprising because, looking back at equation 3.1 and 3.12, which presented kappa and total difference respectively, there seems to be no reason to assume that the results will be exactly the same.

Table 5.1: Overview of representative maps for each similarity metric

Metric	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Kappa	Map 359	Map 456	Map 541	Map 695	Map 1148	Map 1760	Map 1911
Overall Accuracy	Map 10	Map 359	Map 456	Map 541	Map 695	Map 696	Map 1911
Total Difference	Map 359	Map 456	Map 541	Map 695	Map 1148	Map 1760	Map 1911
Total Allocation Difference	Map 263	Map 453	Map 678	Map 1760	Map 1852	Map 1918	-
Simpson's Diversity Index	Map 65	Map 857	Map 1448	Map 1784	Map 1830	-	-
Total Class Area [Corn]	Map 246	Map 296	Map 325	Map 825	Map 1612	-	-
PLAND [Corn]	Map 246	Map 325	Map 564	Map 704	Map 843	-	-
Total Quantity Difference	Map 1057	Map 1472	Map 1833	Map 1978	-	-	-
Shannon's Diversity Index	Map 61	Map 1248	Map 1454	Map 1457	-	-	-
Total Class Area [Residential]	Map 217	Map 332	Map 604	-	-	-	-
PLAND [Residential]	Map 264	Map 1422	Map 1877	-	-	-	-
Total Class Area [Nature]	Map 0	Map 8	Map 35	-	-	-	-
PLAND [Nature]	Map 0	Map 8	Map 35	-	-	-	-
Quantity Difference [Nature]	Map 0	Map 1	Map 8	-	-	-	-

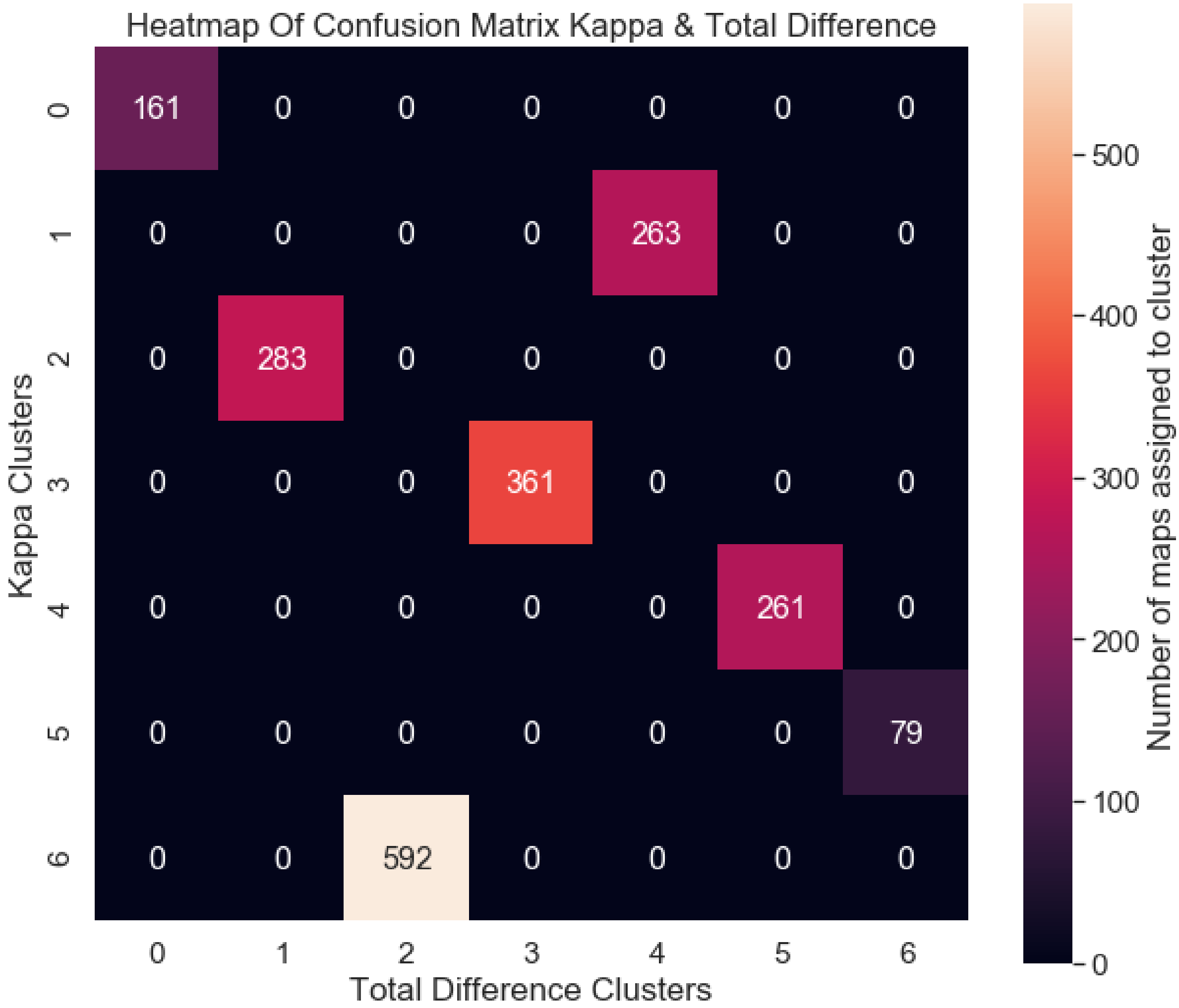
Not only is kappa represented as a normalized fraction, where total difference is calculated as the sum of cells that are considered in disagreement, their mathematical foundation also vary. Where kappa can be simplified as the observed accuracy corrected by the random chance of assigning cells a given value, knowing the underlying data, total difference is a summation of the different types of disagreement, namely quantity and allocation disagreement. Although all other metric pairs in this research suggest otherwise, kappa and total difference show that seemingly different metrics can lead to exactly the same results when clustering.

5.3 Kappa and Overall Accuracy

Kappa and overall accuracy showed similar results in the Chapter 4. The confusion matrix comparing the map allocation to clusters, as visible in Table 5.2, highlights this similarity in results while also confirming that there is a difference between the two. This difference is found in 86 maps from kappa's cluster 3 being assigned to overall accuracy's cluster 6. Given how overall accuracy is a component used in calculating kappa, some similarity in results was not unexpected. According to the clustering, one might assume that there is little difference between clustering on either of these two metrics.

To find out the significance of this difference in cluster allocation, the CART trees for kappa and overall accuracy are compared. The results of the attribute comparisons are presented in Table 5.2. According to these values the tree for kappa seems to be simpler. With only six and seven variables used for either metric, and three of those being unique it is likely that the small differences found in cluster allocation does not necessarily mean that the resulting CART output is very similar.

The CART trees are presented in Figure 5.3 and Figure 5.4 for respectively kappa and overall accuracy. Blue nodes in the tree are decision nodes where maps either follow the input variable or they do not. The white nodes are the leaf nodes and represent the end of a branch. While some similarity is seen between the two trees, there is not a single branch with no divergence. One of the key similarities between the trees is how most of cluster 7 for both metrics ends up in the rightmost leaf node. In all other branches



	0	1	2	3	4	5	6
0	161	0	0	0	0	0	0
1	0	0	0	0	263	0	0
2	0	283	0	0	0	0	0
3	0	0	0	361	0	0	0
4	0	0	0	0	0	261	0
5	0	0	0	0	0	0	79
6	0	0	592	0	0	0	0

Figure 5.1: Confusion matrix of map allocation to clusters for kappa and total difference

the resulting maps are harder to trace. This implies that kappa and overall accuracy are not as similar as the representative maps and the allocation matrices might suggest.

Table 5.2: Comparison of kappa and overall accuracy CART trees on four basic tree attributes.

Metric	**Tree Depth**	**Total Decision Nodes**	**Number of Variables Used (% of total variables)**	**Number of Different Variables (% of used variables)**
Kappa	5	9	6 (24%)	3 (50%)
Overall Accuracy	6	11	8 (32%)	3 (37,5%)

5.4 Categorical Metric Comparison

Noticeable differences were also observed for the categorical metrics percentage landscape, total class area, and quantity difference. Table 5.3 provides an overview of the tree comparisons for the seven potential outcomes. The category residential leads to exactly the same CART trees. This is in line with the representative maps presented in section 5.1, and the confusion matrices for cluster allocation on these pairs which can be found in Appendix E. What is also of note is that for both the residential and nature land-use categories all five of these cases have a tree depth of only two, thus splitting the data on only two out of twenty five available variables. This is also different from the corn category and the two metrics that were presented in the previous section. A possible explanation for this lack of depth is that, especially when clustering on a specific land-use category, there must be enough variance in the

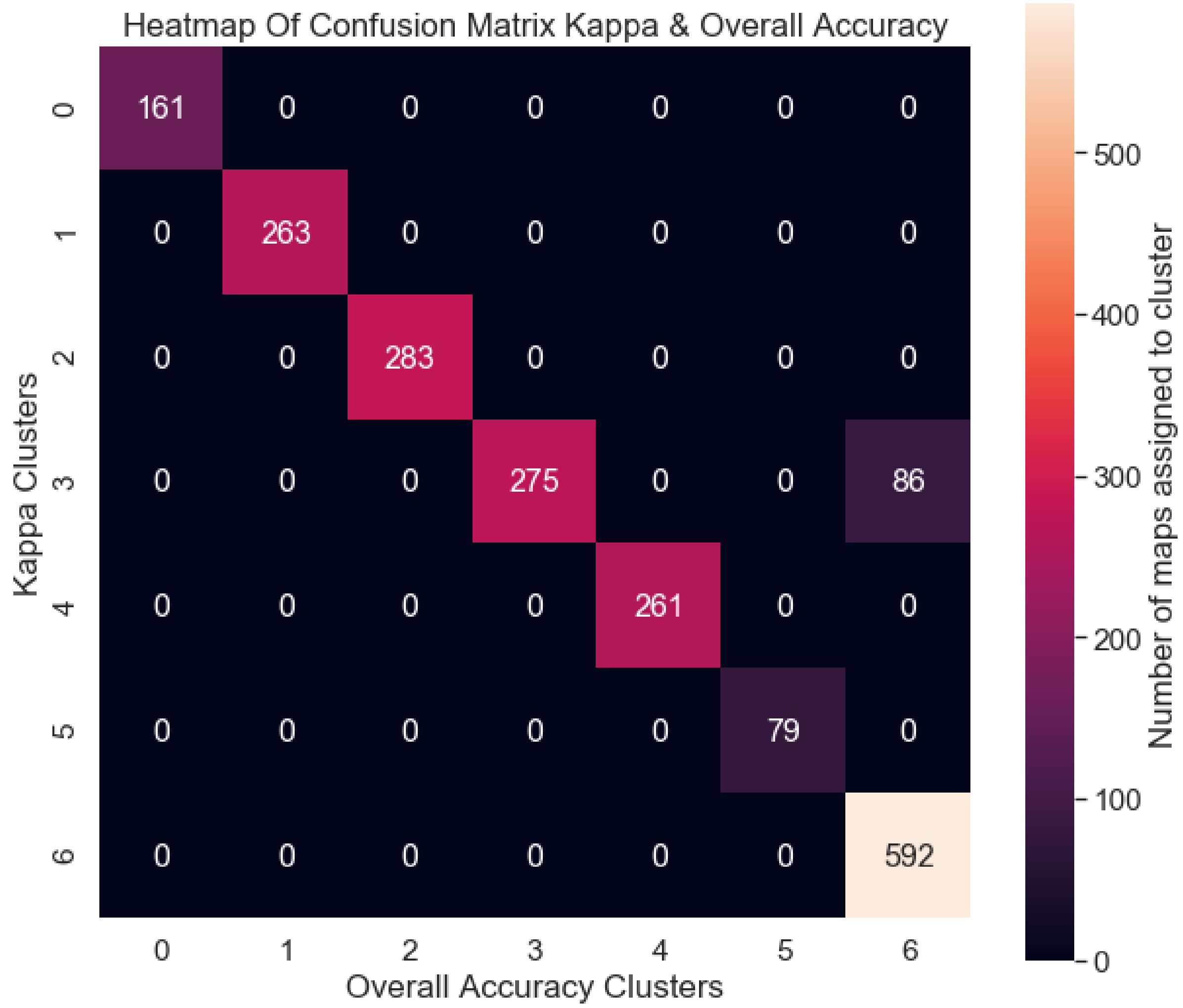


Figure 5.2: Confusion matrix of map allocation to clusters for kappa and overall accuracy

outcomes of that land-use category across all the maps for clustering, and eventual CART analysis, to show diverging results. Across 2,000 maps, residential land-use saw only 15 different results in total land-use allocation for that category. For nature this was 737 and corn, with its higher tree depth, had 1686 different land-use allocations.

An important difference between the residential and nature land-use categories is that even with the small trees, a difference is still observed. While percent landscape and quantity difference have exactly the same CART tree, total class area splits the tree based on a different variable first. With only two decision nodes, having one of the two being a different variable is very impactful, while they also split the data on the variable they agree on in different steps. It is noteworthy that quantity difference and percentage landscape are the same since in this example quantity difference is calculated with absolute land values whereas percentage landscape is normalized. Since total class area also works with absolute values it would be expected that it would show full agreement with quantity difference instead. This is likely in part because of the way quantity difference is calculated, where it does not consider components of disagreement related to allocation. Total class area considers the absolute difference in land-use category assignment for two maps, whereas quantity difference would not consider some of these as they are part of the shift component of allocation disagreement (Pontius & Santacruz, 2014).

It is however just as interesting that percentage landscape and total class area calculated for the corn category lead to seemingly different CART trees. Since percentage landscape is total class area normalized for total land area it is relevant to find that a difference can be found when clustering on normalized values. Percentage landscape has a larger tree, but most notably the overlap in variables used is larger

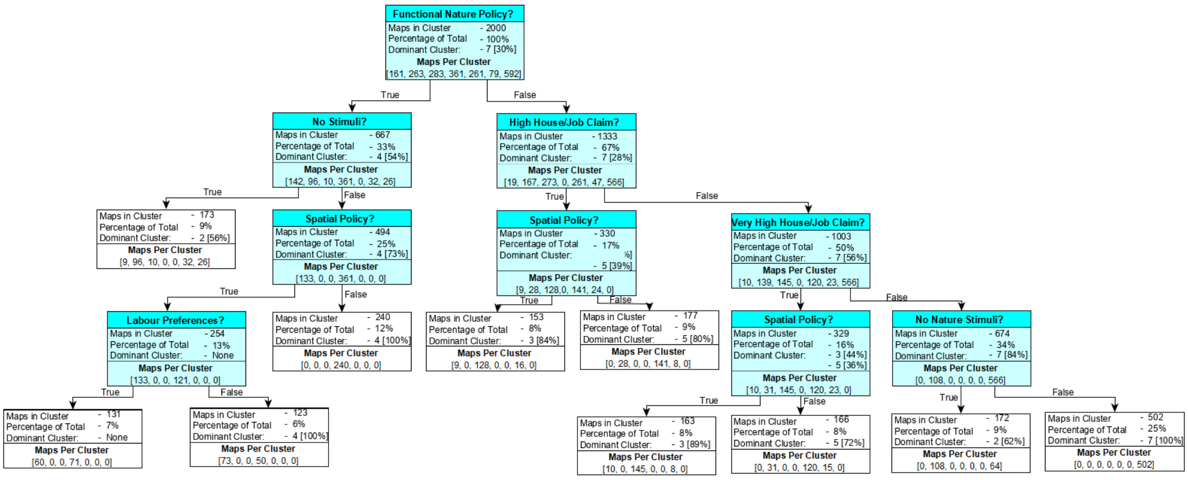


Figure 5.3: CART tree for kappa clusters

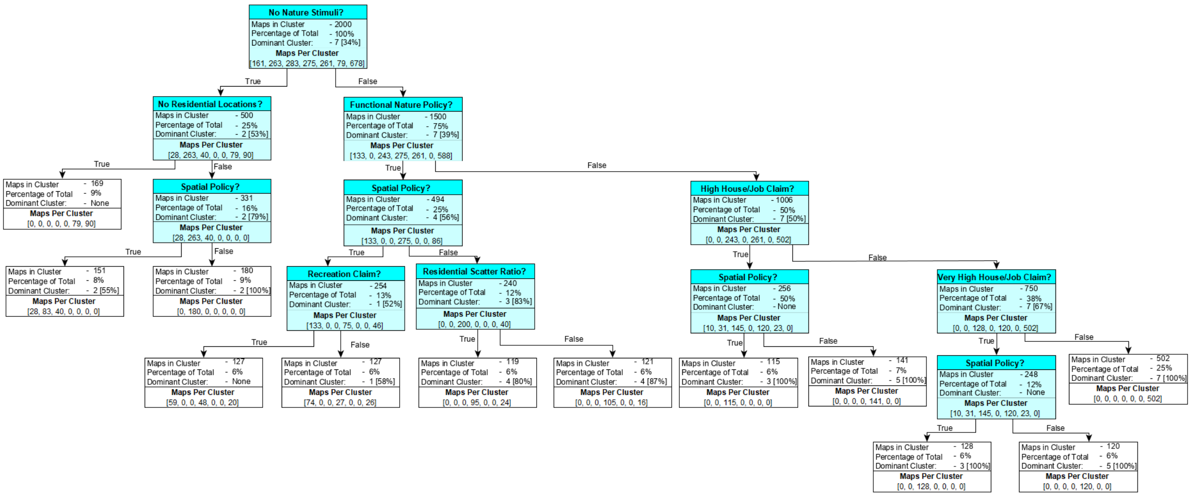


Figure 5.4: CART tree for overall accuracy clusters

than the comparison presented for kappa and overall accuracy in section 5.3. This is perhaps the most relevant aspect of Table 5.3 as it highlights a focus on different scenarios as each variable can be considered as a lever that can be turned on and off, either allowing or preventing certain developments from happening in the Land Use Scanner model.

5.5 CART Attribute Overview

A global picture of the different CART tree results is formed by looking at them side by side. This overview is presented in Table 5.4. Across all variables different values are observed. The variance in the total number of decision nodes is of interest as it shows how different metrics are able to section the data differently. Although it is not possible to say that more is better, it is of interest that the two ’outsider’ metrics that calculate diversity have the most decision nodes. This does come with the highest percentage of duplicate nodes in the tree as well.

The number of variables used shows that clustering on a categorical metric does not necessarily mean that the sectioning of the data is focused merely on that metric’s land-use category. Total class area and percentage landscape for the corn category both show the higher end of the number of variables used, while being a categorical metric. With the simple CART trees for the residential and nature categories

Table 5.3: Overview of the attributes of the CART trees for categorical metrics.

Metric	Tree Decision Depth	Total Decision Nodes	Number of Variables Used (% of total variables)	Number of Different Variables Used (% of used variables)
TCA [Residential]	2	2	2 (8%)	0 (0%)
PLAND [Residential]	2	2	2 (8%)	0 (0%)
TCA [Corn]	5	11	6 (24%)	4 (67%)
PLAND[Corn]	6	14	8 (32%)	4 (50%)
TCA [Nature]	2	2	2 (8%)	1 (50%)
PLAND [Nature]	2	2	2 (8%)	1 (50%)
Quantity Difference [Nature]	2	2	2 (8%)	1 (50%)

it is hard to draw proper conclusions without expanding to a larger sample set of categories.

Table 5.4: Overview of the attributes of the CART trees for all metrics.

Metric	Tree Decision Depth	Total Decision Nodes	Number of Variables Used (% of total variables)	Number of Duplicate Entries (% of decision nodes)
Kappa	4	9	6 (24%)	3 (33%)
Overall Accuracy	5	11	8 (32%)	3 (27%)
Total Difference	4	9	6 (24%)	3 (33%)
Total Allocation Difference	3	5	3 (12%)	2 (40%)
Total Quantity Difference	4	8	8 (32%)	0 (0%)
Shannon's Diversity Index	5	15	6 (24%)	8 (53%)
Simpson's Diversity Index	5	14	7 (28%)	7 (50%)
TCA [Residential]	2	2	2 (8%)	0 (0%)
PLAND [Residential]	2	2	2 (8%)	0 (0%)
TCA [Corn]	5	11	6 (24%)	4 (36%)
PLAND[Corn]	6	14	8 (32%)	6 (43%)
TCA [Nature]	2	2	2 (8%)	0 (0%)
PLAND [Nature]	2	2	2 (8%)	0 (0%)
Quantity Difference [Nature]	2	2	2 (8%)	0 (0%)

5.6 Scenario Defining Input Variables

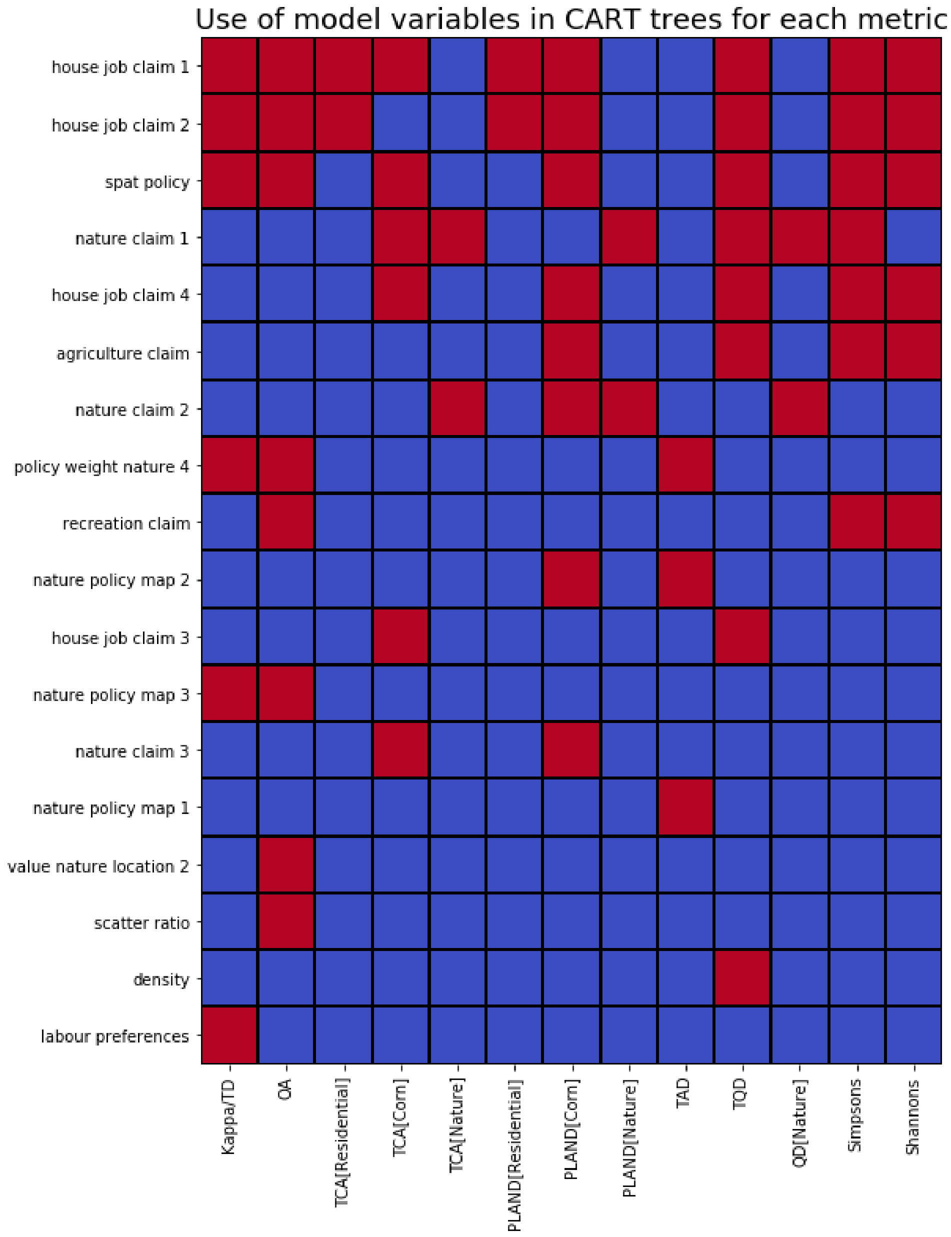
A further way of highlighting the differences between metrics is to visualize the difference in Land Use Scanner model variables used. Figure 5.5 shows a red cell when a metric uses a specific variable, and a blue cell if it does not. Only 18 out of 25 variables are shown as the remaining 7 were not used in any CART tree. Each variable allows for certain aspects of land-use change to be more likely when they are used, and thus lead to different land-use maps. Often times these variables may be linked to specific policies.

It is apparent that three of the house job claims, spatial policy and nature claim 1 are more used between the metrics. Besides the common use of some variables, the majority of rows in the figure also show variables that are used by between one to three metrics. This suggests that, while there are shared variables in the pool, different metrics are able to extract different information from the data.

5.7 Chapter Conclusion

This chapter expanded on the individual metric analysis in chapter 4 by comparing results cross metric instead. The representative maps found for each metric's cluster suggested that there might be a difference in clustering results between most metrics. This was largely confirmed by computing the confusion matrices showing the allocation of maps to clusters between metrics. The two notable exceptions to this were the pairs of kappa and total difference, and kappa with overall accuracy. kappa and total difference stood out because they lead to the exact same results when used for clustering maps, including the CART analysis. Kappa and overall accuracy showed a lot of similarity for cluster allocation, only disagreeing on the assignment of 86 out of 2,000 maps. Further CART analysis did show that this seemingly small difference in cluster allocation does lead to observable differences in the resulting decision trees.

Figure 5.5: Confusion matrix of variables used for each metric (red=yes, blue=no)



A comparison of the percentage landscape and total class area metrics for the land-use categories residential, corn, and nature showed simplistic tree outcomes for both the residential and nature land-use categories. This might be due to the lack of outcome variance for these land-use categories in the Land Use Scanner model. On the other hand, the corn category showed two different trees for percentage landscape and total class area. Of special interest for this is the final column of the table, which highlights the difference in variables used during the CART analysis to split the data.

One of the reasons to select both total class area and percentage landscape, was to find out whether a normalized version of a metric might lead to different results. Although the similarity and peculiar results of the residential and nature runs are not fully explained, the corn category alone indicates that scenario discovery results can definitely vary between normalized and non-normalized metrics.

Comparisons between most metrics show largely diverging results. Since this is expected it was not presented, but it is important to note. If each categorical metric with its own land-use category is considered a unique metric, this research compares a total of 14 metrics. This means a total of 91 metric comparisons were made. Besides the few overlaps presented earlier, all other comparisons show large differences for both the confusion matrices as well as the CART trees. This difference in results between most metrics was also found in the global comparison of CART trees where most metrics had varying tree attribute parameters (Table 5.4) and, although some model variables were commonly shared, there was also a difference in input variables used between metrics (Figure 5.5).

Chapter 6

Discussion

As the application of scenario discovery to land-use change models is a novel concept it is not surprising that new challenges were found, while some approaches might not have proven as effective as hoped. This chapter highlights the main discussion points that were encountered during the research.

During the clustering step of the analysis only hierarchical clustering algorithm was used, with a fixed set of parameters. Instead of the hierarchical clustering algorithm other alternatives should be tested to better understand the impact of choices made. This also includes the number of clusters selected for final clustering. Although a methodical approach was used to decide the final number of clusters it was nonetheless a subjective choice. Furthermore, it is unclear whether conventional approaches for optimal cluster numbers apply in this research. Analysis performed to obtain the cluster numbers showed that cluster numbers are largely about making trade-offs. Experimenting with the cluster numbers will provide better insight into what approach is desired when clustering on similarity metrics during the scenario discovery process.

Initial results in this research showed that a lack of information for the clustering algorithm might lead to similar results. This was apparent for the categorical metrics, where the outcome variance across land-use categories was intentionally different for the three selected categories. If this is true, a simple analysis of the large set of generated land-use maps can provide information as to whether further scenario discovery steps are worth pursuing. If, for example, it is decided that a categorical similarity metric is required for clustering, but that the categories of interest are not showing the degree of outcome variance which is required for clustering, then further steps would provide no value.

Representative maps played a key role for comparing the clustering results for each metric. While they are an insightful tool, especially considering their ease of use, it is unclear how reliable this approach is. Due to the nature of clustering it is challenging to efficiently analyze what a cluster contains. This is no different with regards to representative maps. As such it is unclear how representative a representative map truly is for a given cluster. Additionally, the CART analysis showed that a given cluster could often end up on different leaves of the tree. This suggests that a single representative map approach is either not sufficient, or that the number of clusters might not be enough.

While no conclusive framework is achieved as to what similarity metric should be used in what situation, it is clear that different similarity metrics will generally lead to different results, unless the metrics are inherently similar. Based on current work, the best approach for similarity metric selection is to link it to the modelling objective. If the model needs to answer a policy objective related to urbanization, then selecting a metric that incorporates this makes more sense. However, considering the novelty of applying scenario discovery to land-use change models, it is likely prudent to not be satisfied with selecting only one similarity metric for the job, and instead selecting more as time allows. Not only does this help manage the lack of clarity on the effects of similarity metrics on scenario discovery results, it might simply be preferable in general to use multiple similarity metrics to study the data in a different context.

Chapter 7

Conclusion and Future Research

Land-use change modelling is a tool used for understanding land-use change and to advise policy makers. As these models represent complex systems they inherently contain uncertain aspects. On top of this, the systems they model contain uncertain factors. Contrary to standard land-use change modelling approaches, these uncertain factors can be studied and incorporated with the use of scenario discovery. scenario discovery is an approach that allows for processing of large sets of output. In the case of land-use change models, this output is maps. Because maps are an unusual output format to be explored with scenario discovery, further investigation is required. One of these aspects requiring further exploration is the use of similarity metrics for calculating a relevant difference between maps. This research sets the first step to better understanding the impact that different similarity metrics have on later scenario discovery steps.

The main findings of this research can be summarized as follows:

- Of the 91 metric pairs compared, only 6 showed similar or the same results
- Metrics with a different mathematical approach can still lead to the same or similar results (i.e. kappa and total difference)
- Similarity in cluster results may still lead to noticeable differences in later scenario discovery steps
- It seems plausible that similarity metrics can be classified as being usable for a specific purpose
- If there is a lack of outcome variance (information) in land-use change modelling results, metrics may provide similar results
- Normalizing a metric does not necessarily provide the same results as its non-normalized version
- A noticeable difference was observed between similarity metrics in the use of input variables for their respective CART trees

The next section answers the research questions that led to the former findings. Following that, avenues for future research are presented. The chapter concludes with a reflection on the contributions of this research.

7.1 Revisiting the Research Questions

The research questions leading to the conclusions presented previously will be answered next. This research aimed to answer one main research question, which was split into four sub questions.

Sub Research Question 1

What similarity metrics exist for comparing maps?

The first sub research question was answered through a literature search. Applications of land-use change modelling were reviewed to find commonly used similarity metrics for comparing maps. It is apparent that the last two decades saw an increase in the number of similarity available. The metrics of interest can be split into two categories: composition-based and those focusing on spatial configuration. Composition-based metrics are those that focus on aspects that are easy to quantify, such as the number of cells of a specific land-use category. Spatial configuration metrics are generally more difficult to calculate as they focus on relative positioning of patches. As metrics become more difficult to calculate, their interpretation value as a basis for clustering is also pulled into question. However, commonly used metrics that are relative straight forward do exist.

Kappa is a traditional, and frequently used metric across multiple research fields and is still used to this day. It has seen numerous additions to it over the years to tackle some of its weaknesses. Nonetheless, there are also strong voices against the use of this metric and its adapted versions. As such alternatives are recommended to be used over kappa (Pontius Jr & Millones, 2011).

Beyond this dichotomy, a metric attribute of interest is whether a metric is categorical in nature or not. These metrics focus their calculation on a specific land-use category in the model. However, these metrics are also available in a variety of different flavors leading to different results. It is likely that different policy questions will require different metrics, but it is too early to make any conclusive statements on what metrics should and should not be used.

Sub Research Question 2

How can similarity metrics be embedded in an agglomerative mode clustering algorithm?

Clustering using a similarity metric largely follows the same procedure as for any other distance value. One key consideration that must be made is whether the similarity metric's output can be used as distance value for clustering. First, for metrics that are normalized between zero and one, the metric is likely to measure similarity. In such cases it makes more sense to cluster on the complement of the metric (1 - metric value) so that clustering happens on lowest disagreement first, instead of the highest disagreement. Second, metrics that calculate a unique value per map require all combinations of map pairs to be calculated manually instead. In such cases it is sensible to consider what subtracting two of these map values from one another to obtain a distance value signifies. For a metric such as total class area this difference makes intuitive sense as it signifies the absolute difference in land-use allocation between two maps for a chosen class. For other metrics, such as Simpson's diversity index, clumpiness, or fractal dimension such a subtraction between values does not intuitively provide meaning to the value.

Sub Research Question 3

What is the effect of the different similarity metrics on the resulting clusters of maps?

Different similarity metric have shown to lead to varying clustering results. First, With 14 different metrics and a total of 91 different comparisons between metrics, only 6 combinations showed the same, or similar, results. Four of these were for categorical metrics analysing the same land-use category. For these cases it is also unclear to what degree this similarity in results stems from a lack of outcome variance for those land-use categories in the land-use change model.

Second, a considerable difference between metrics is also found in the number of recommended clusters. Varying between three to seven, allocating the set of maps across a different number of clusters is always going to lead to differences.

Finally, comparing the representative maps. it seems plausible that each metric defined their own unique clusters focusing on different land-use allocations. However, as the representative map is only that map that is considered on average the most similar to all other maps in the cluster, it provides only an initial glimpse into what the clusters contain. Nonetheless, the initial results showed that, while a few land-use categories were highly represented among all similarity metrics, all metrics managed to highlight land-use change narratives for their set of representative maps. More importantly, result showed different similarity metrics are able to split the data in a manner that is useful for later analysis. Two examples are Shannon's diversity index seemed to create a cluster of extreme values, whereas total allocation difference was able to pick up on relocation on nature land-use that other similarity metrics seemed to miss.

Sub Research Question 4

What is the effect of the different similarity metrics
on the resulting scenarios discovered through CART?

Similarly to the clustering results, most CART trees showed different results. Although a systematic, and widely agreed upon, method of comparing decision trees is not available, a simple analysis using tree attributes showed that after applying the CART analysis only three out of six results that were similar after clustering also lead to the exact same CART tree. These are the cases where the clustering results were exactly the same to begin with. The remaining three cases instead showed relatively large differences in CART trees.

For the three categorical metrics focusing on the land-use category nature, after clustering results seemed to suggests that cluster allocation of maps was exactly the same, CART analysis showed that this might have been deceptive as one of the metrics had a different CART tree. These metrics only had two decision nodes, but they disagreed on one of the variables used. This highlights that care must be taken to not draw conclusions based on seemingly overlapping cluster allocation of maps between different metrics, as CART analysis may highlight that there are subtle differences.

The choice of metric does affect model input variables identified as important through CART analysis. Out of 25 different input variables, 18 were used in at least one CART tree. While some of use saw frequent use for multiple metrics, a majority of these variables were only used for one to three metrics. This highlights how different metrics extract different information from the map set.

Main Research Question

How does the choice of similarity metric affect Scenario
Discovery results of land-use change modelling?

Different similarity metrics seem to highlight different aspects of Land Use Scanner output. Since the CART analysis showed a difference in variable usage it is likely that a different scenario narrative would result when selecting a different metric. As there is no point of reference, it is unclear to what degree this is tied to the similarity metrics that were selected or the peculiarities of the chosen land-use change model. However, given the fourteen metrics that were analysed in this research, it is not possible to make conclusive statements which of these are suitable for use with a given policy question. The categorical metrics with simplistic CART trees are likely not suitable, but that still leaves nine metrics to choose from. Although it is apparent that different similarity metrics lead to different results, what these results entail for different metrics is not clear. Whether or not all these varying results are valuable is unclear without a better understanding of the clustering results.

With that said, future application of scenario discovery to land-use change models should consider using multiple similarity metrics. If anything, this research has shown that different similarity metrics lead to different results. At best, it has shown that similarity metrics can be applied with a specific purpose in mind. Until this latter point is better understood, the usage of multiple similarity metrics allows for exploring model results using different mathematical perspectives.

7.2 Future Research

During the research various aspects were found that warrant further attention. The following topics are the most notable areas for future research.

Similarity metrics play a crucial role in this research as the tool used for comparing maps, however only a small number were tested. Not only are there numerous metrics out there that are worth comparing, the next step after clustering on a metric must also be explored. If we cluster on a given metric, do the resulting clusters partition the data in a way that we are interested in, or are we always none the wiser with what the results might be? This is of specific interest when a model is created, and analyzed, for a clear purpose. In such a case it must not only be apparent what a similarity does for the resulting clusters, it can also be of value to apply multiple similarity metrics. However, should these be applied in parallel, as different metrics were in this research, or would an aggregation, perhaps some kind of multi-criteria approach, be a better solution? Or would this lead to the same issues that a binary threshold has in a multi-stakeholder situation?

Chapter 4 introduced the real possibility that a framework for the selection of similarity metrics for scenario discovery on land-use change models is feasible. However, the approach followed in this research is far from conclusive to prove this point, and only hints at the possible attributes of similarity metrics and how they compare maps. Research is required that tests similarity metrics on different models to truly detect patterns and transferability of results.

A distinction where understanding cluster results better is of interest is in the case of normalized versus non-normalized metrics. In this research total class area is a non-normalized metric, whereas percentage landscape is a normalized version of total class area. Results showed that normalized and non-normalized metrics do not necessarily provide the same results. Without a better understanding of what the clusters contain, it is unclear in which situations a normalized or non-normalized metric would lead to clusters that are able to provide robust answers to policy questions. Note that any clustering result is likely able to provide answers to the questions that are posed. We should also consider whether changing the distance value, by using a different similarity metric, might lead to clusters that are better able to extract the information that the model can provide. Thus providing a more robust answer. The dictated policy question changing the similarity metric that is best able to extract the information is the challenge that must be tackled.

7.3 Scientific Impact

While scenario discovery as a method is gaining traction, the use of multiclass scenario discovery is still less common. This research has shown how multiclass scenario discovery can be used to engage with a new domain. In a more general sense, it provides another example of how multiclass classification of results is a viable, and likely preferable, alternative to using the standard binary classification approach. While not the core focus of this research, a classification approach for scenario discovery that discards large parts of its data seems counterproductive. One of the arguments to use scenario discovery over other approaches, such as story and simulation, is after all that it deals with uncertainty better by incorporating a larger set of scenarios.

7.4 Societal Impact

Land-use change models already see use in policy advice. With scenario discovery this adds the possibility for the models to be more robust in an uncertain world. Furthermore, the steps of scenario discovery can also help improve land-use change models to begin with. Although scenario discovery in this context needs a lot of work, if knowledge gaps are filled, such as what similarity metrics should be used in what case, and for which models, with a better understanding of what different clustering algorithms and their parameters have for effects on the results, land-use change models can be scrutinized in a way never done before.

This research has helped progress the application of scenario discovery to land-use change models by exploring the effect that similarity metric selection has on the clustering step and later scenario discovery results. While some results are intuitive, such as different similarity metrics often leading to different results, some unexpected outcomes were also observed. Furthermore, an initial glimpse was offered into how different similarity metrics can be selected to extract different information from model results.

References

Aguejdad, R., Houet, T., & Hubert-Moy, L. (2017). Spatial validation of land use change models using multiple assessment techniques: A case study of transition potential models. *Environmental Modeling & Assessment*, *22*(6), 591–606.

Alcamo, J. (2008). Chapter six the sas approach: Combining qualitative and quantitative knowledge in environmental scenarios. In J. Alcamo (Ed.), *Environmental futures* (Vol. 2, p. 123 - 150). Elsevier. Retrieved from `http://www.sciencedirect.com/science/article/pii/S1574101X08004067` doi: https://doi.org/10.1016/S1574-101X(08)00406-7

Al-Shalabi, M., Billa, L., Pradhan, B., Mansor, S., & Al-sharif, A. (2012, 09). Modelling urban growth evolution and land-use changes using gis based cellular automata and sleuth models: The case of sana'a metropolitan city, yemen. *Environmental Earth Sciences*, *70*. doi: 10.1007/s12665-012-2137-6

Argent, R. M., Sojda, R. S., Giupponi, C., McIntosh, B., Voinov, A. A., & Maier, H. R. (2016). Best practices for conceptual modelling in environmental planning and management. *Environmental Modelling & Software*, *80*, 113 - 121. Retrieved from `http://www.sciencedirect.com/science/article/pii/S1364815216300433` doi: https://doi.org/10.1016/j.envsoft.2016.02.023

Ash, E., Macdonald, D., Cushman, S., Noochdumrong, A., Redford, T., & Kaszta, (2021, 01). Optimization of spatial scale, but not functional shape, affects the performance of habitat suitability models: a case study of tigers (panthera tigris) in thailand. *Landscape Ecology*, 1-20. doi: 10.1007/s10980-020-01105-6

Bibi, F., Ali, Z., et al. (2013). Measurement of diversity indices of avian communities at taunsa barrage wildlife sanctuary, pakistan. *The Journal of Animal & Plant Sciences*, *23*(2), 469–474.

Borsboom, J., Regt, W., & Schotten, K. (2002). Land use scanner: the continuous cycle of application, evaluation and amelioration in landuse modelling. *European Regional Science Association, ERSA conference papers*.

Bošnjak, L., Karakatič, S., & Podgorelec, V. (2015). Using similarity-based selection in evolutionary design of decision trees. In *2015 38th international convention on information and communication technology, electronics and microelectronics (mipro)* (pp. 1206–1211).

Bradfield, R., Wright, G., Burt, G., Cairns, G., & Van Der Heijden, K. (2005). The origins and evolution of scenario techniques in long range business planning. *Futures*, *37*(8), 795 - 812. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0016328705000042` doi: https://doi.org/10.1016/j.futures.2005.01.003

Bryant, B. P., & Lempert, R. J. (2010). Thinking inside the box: A participatory, computer-assisted approach to scenario discovery. *Technological Forecasting and Social Change*, *77*(1), 34 - 49. Retrieved from `http://www.sciencedirect.com/science/article/pii/S004016250900105X` doi: https://doi.org/10.1016/j.techfore.2009.08.002

Celio, E., Koellner, T., & Grêt-Regamey, A. (2014). Modeling land use decisions with bayesian networks: Spatially explicit analysis of driving forces on land use change. *Environmental Modelling & Software*, *52*, 222 - 233. Retrieved from `http://www.sciencedirect.com/science/article/pii/S1364815213002570` doi: https://doi.org/10.1016/j.envsoft.2013.10.014

Chaudhuri, G., & Clarke, K. C. (2014). Temporal accuracy in urban growth forecasting: A study using the sleuth model. *Transactions in GIS*, *18*(2), 302–320.

Cimatti, M., Ranc, N., Benítez-López, A., Maiorano, L., Boitani, L., Cagnacci, F., ... Santini, L. (2021, 01). Large carnivore expansion in europe is associated with human population density and land cover changes. *Diversity and Distributions*. doi: 10.1111/ddi.13219

Claassens, J., Koomen, E., & Rijken, B. (2017, 12). Actualisering landgebruiksimulatie deltascenario's achtergronddocument bij ruimtescanner inzet. doi: 10.13140/RG.2.2.19038.59209

Cox, M. (2020). *Scenario discovery in land use change models* (Master's Thesis, Delft, University of Technology). Retrieved from https://repository.tudelft.nl/islandora/object/uuid:2ec47289-dd3d-49a3-a1c8-addcdff405bf

Dalla-Nora, E. L., de Aguiar, A. P. D., Lapola, D. M., & Woltjer, G. (2014). Why have land use change models for the amazon failed to capture the amount of deforestation over the last decade? *Land Use Policy*, *39*, 403 - 411. Retrieved from http://www.sciencedirect.com/science/article/pii/S0264837714000246 doi: https://doi.org/10.1016/j.landusepol.2014.02.004

Gaudreau, J., Perez, L., & Drapeau, P. (2016). Borealfiresim: A gis-based cellular automata model of wildfires for the boreal forest of quebec in a climate change paradigm. *Ecological Informatics*, *32*, 12–27.

Gerst, M. D., Wang, P., & Borsuk, M. E. (2013). Discovering plausible energy and economic futures under global change using multidimensional scenario discovery. *Environmental modelling and software*, *44*, 76–86.

Hagen-Zanker, A. (2006). Comparing continuous valued raster data: A cross disciplinary literature scan.

Halim, R. A., Kwakkel, J. H., & Tavasszy, L. A. (2016). A scenario discovery study of the impact of uncertainties in the global container transport system on european ports. *Futures*, *81*, 148 - 160. Retrieved from http://www.sciencedirect.com/science/article/pii/S0016328715001342 (Modelling and Simulation in Futures Studies) doi: https://doi.org/10.1016/j.futures.2015.09.004

Harvey, E. P., Cardwell, R. C., McDonald, G. W., van Delden, H., Vanhout, R., Smith, N. J., ... van den Belt, M. (2019). Developing integrated models by coupling together existing models; land use, economics, demographics and transport in wellington, new zealand. *Computers, Environment and Urban Systems*, *74*, 100–113.

Hu, M., Wang, X., Wen, X., & Xia, Y. (2012). Microbial community structures in different wastewater treatment plants as revealed by 454-pyrosequencing analysis. *Bioresource technology*, *117*, 72–79.

Jin, H., & Mountrakis, G. (2013). Integration of urban growth modelling products with image-based urban change analysis. *International Journal of Remote Sensing*, *34*(15), 5468-5486. Retrieved from https://doi.org/10.1080/01431161.2013.791760 doi: 10.1080/01431161.2013.791760

Kityuttachai, K., Tripathi, N., Tipdecho, T., & Shrestha, R. (2013, 04). Ca-markov analysis of constrained coastal urban growth modeling: Hua hin seaside city, thailand. *Sustainability*, *5(4)*, 1480-1500. doi: 10.3390/su5041480

Kwakkel, J. (2019, 02). A generalized many-objective optimization approach for scenario discovery. *Futures & Foresight Science*, *1*. doi: 10.1002/ffo2.8

Kwakkel, J. H. (2017). The exploratory modeling workbench: An open source toolkit for exploratory modeling, scenario discovery, and (multi-objective) robust decision making. *Environmental Modelling & Software*, *96*, 239 - 250. Retrieved from http://www.sciencedirect.com/science/article/pii/S1364815217301251 doi: https://doi.org/10.1016/j.envsoft.2017.06.054

Lempert, R., Bryant, B., & Bankes, S. (2008, 01). Comparing algorithms for scenario discovery.

Liu, D., Zheng, X., & Wang, H. (2020). Land-use simulation and decision-support system (landsds): Seamlessly integrating system dynamics, agent-based model, and cellular automata. *Ecological Modelling*, *417*, 108924. Retrieved from http://www.sciencedirect.com/science/article/pii/S0304380019304326 doi: https://doi.org/10.1016/j.ecolmodel.2019.108924

McGarigal, K. (2015). FRAGSTATS HELP. (https://www.umass.edu/landeco/research/fragstats/documents/fragstats.help.4.2.pdf [Accessed: 27-10-2020])

McJeon, H. C., Clarke, L., Kyle, P., Wise, M., Hackbarth, A., Bryant, B. P., & Lempert, R. J. (2011). Technology interactions among low-carbon energy technologies: What can we learn from a large number of scenarios? *Energy Economics*, *33*(4), 619 - 631. Retrieved from http://www.sciencedirect.com/science/article/pii/S0140988310001866 (Special Issue on The Economics of Technologies to Combat Global Warming) doi: https://doi.org/10.1016/j.eneco.2010.10.007

Nagabhatla, N., Finlayson, M., & Senaratna Sellamuttu, S. (2012, 06). Assessment and change analyses (1987-2002) for tropical wetland ecosystem using earth observation and socioeconomic data. *European Journal of Remote Sensing*, *45*, 215-232. doi: 10.5721/EuJRS20124520

Nussbaumer, S., Huggel, C., Schaub, Y., & Walz, A. (2013, 08). Local land-use change based risk estimation for future glacier lake outburst flood. *Natural Hazards and Earth System Sciences Discussions*, *1*, 4349-4387. doi: 10.5194/nhessd-1-4349-2013

Pontius, R. G., & Santacruz, A. (2014). Quantity, exchange, and shift components of difference in a square contingency table. *International Journal of Remote Sensing*, *35*(21), 7543-7554. doi: 10.1080/2150704X.2014.969814

Pontius Jr, R. G., & Millones, M. (2011). Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, *32*(15), 4407–4429.

Punia, A., Joshi, P. K., & Siddaiah, N. S. (2021, 02). Characterizing khetri copper mine environment using geospatial tools. *SN Applied Sciences*, *3*. doi: 10.1007/s42452-021-04183-6

Rounsevell, M., Reginster, I., Araújo, M. B., Carter, T., Dendoncker, N., Ewert, F., ... others (2006). A coherent set of future land use change scenarios for europe. *Agriculture, Ecosystems & Environment*, *114*(1), 57–68.

Rozenberg, J., Guivarch, C., Lempert, R., & Hallegatte, S. (2014). Building ssps for climate policy analysis: a scenario elicitation methodology to map the space of possible future challenges to mitigation and adaptation. *Climatic change*, *122*(3), 509–522.

Schotten, K., Goetgeluk, R., Hilferink, M., Rietveld, P., & Scholten, H. (2001). Residential construction, land use and the environment. simulations for the netherlands using a gis-based land use model. *Environmental Modeling & Assessment*, *6*(2), 133–143.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Simpson, E. H. (1949). Measurement of diversity. *nature*, *163*(4148), 688–688.

SPINLab. (n.d.). *Land use scanner model.* Retrieved from `https://spinlab.vu.nl/research/spatial-analysis-modelling/land-use-scanner-model/`

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote sensing of Environment*, *62*(1), 77–89.

Steinmann, P., Auping, W. L., & Kwakkel, J. H. (2020). Behavior-based scenario discovery using time series clustering. *Technological Forecasting and Social Change*, *156*, 120052.

Tong, S. T., Sun, Y., Ranatunga, T., He, J., & Yang, Y. J. (2012). Predicting plausible impacts of sets of climate and land use change scenarios on water resources. *Applied Geography*, *32*(2), 477 - 489. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0143622811001305` doi: https://doi.org/10.1016/j.apgeog.2011.06.014

Ustaoglu, E., Williams, B., Petrov, L. O., Shahumyan, H., & Van Delden, H. (2018). Developing and assessing alternative land-use scenarios from the moland model: A scenario-based impact analysis approach for the evaluation of rapid rail provisions and urban development in the greater dublin region. *Sustainability*, *10*(1), 61.

van Borsboom, J., de Jong, K., de, N., Nijs, T., Hagen-Zanker, A., CGM, K., & Verburg, P. (2004, 01). The map comparison kit: methods, software and applications.

van Vliet, J., Bregt, A. K., Brown, D. G., van Delden, H., Heckbert, S., & Verburg, P. H. (2016). A review of current calibration and validation practices in land-change modeling. *Environmental Modelling & Software*, *82*, 174–182.

Verburg, P., Schot, P., Dijst, M., & Veldkamp, A. (2004, 01). Land-use change modeling: Current practice and research priorities. *GeoJournal*, *61*, 309-324. doi: 10.1007/s10708-004-4946-y

Yalew, S. G., Mul, M. L., Van Griensven, A., Teferi, E., Priess, J., Schweitzer, C., & van Der Zaag, P. (2016). Land-use change modelling in the upper blue nile basin. *Environments*, *3*(3), 21.

List of Figures

List of Tables

Appendix A

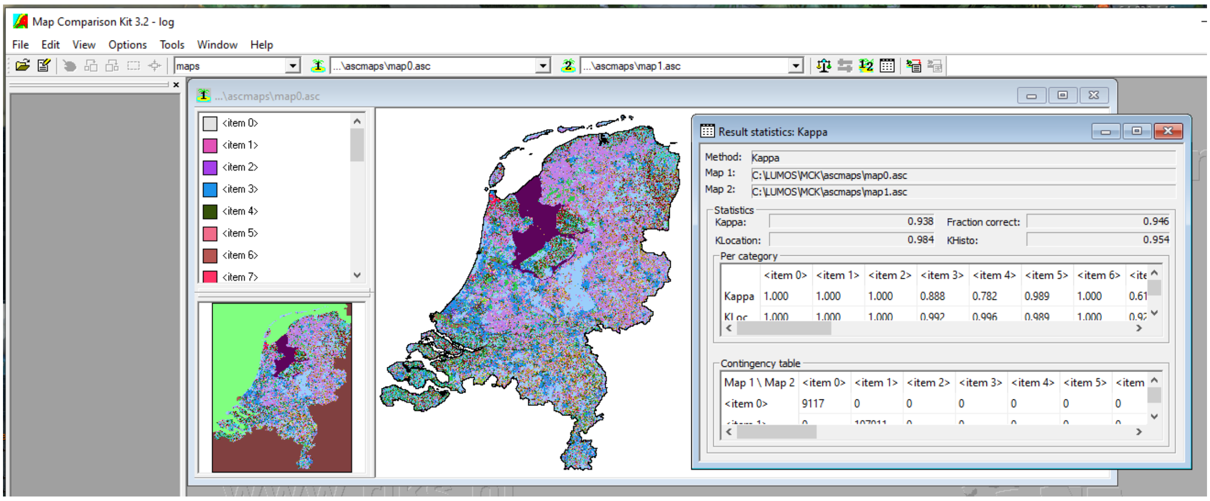
Software and Tools

This research made use of the Map Comparison Kit and Python for calculating map comparison values. Their usage will be discussed in order.

A.1 Map Comparison Kit

The Map Comparison Kit (MCK) is a tool used to compare land-use change maps. It was developed by the Research Institute for Knowledge Systems (RIKS). The available metrics vary from kappa to its deviations, such as fuzzy kappa and fuzzy kappa simulation. It also offers a variety of other options such as patch related metrics, Shannon and Simpson's diversity indices, fractal dimension, and clumpiness. Figure A.1 provides an example Kappa calculation using the Map Comparison Kit for two maps generated with the Land Use Scanner.

While the program is most effective when comparing small sets of maps, in this research it has shown applicable for running large scale comparisons as well. By generating the comparison files that the MCK requires, command prompt can be used to queue up large sets of map comparisons. The advantage of this approach is that the MCK provides reliable output, whereas new implementations would require extra validation. Since most similarity metric options are tied into GIS software, this proved to be a helpful tool.



Result statistics: Kappa

Method: Kappa
Map 1: C:\LUMOS\MCK\ascmaps\map0.asc
Map 2: C:\LUMOS\MCK\ascmaps\map1.asc

Statistics
Kappa: 0.938 Fraction correct: 0.946
KLocation: 0.984 KHisto: 0.954

Per category

	<item 0>	<item 1>	<item 2>	<item 3>	<item 4>	<item 5>	<item 6>	<ite
Kappa	1.000	1.000	1.000	0.888	0.782	0.989	1.000	0.61
KLoc	1.000	1.000	1.000	0.992	0.996	0.989	1.000	0.9?

Contingency table

Map 1 \ Map 2	<item 0>	<item 1>	<item 2>	<item 3>	<item 4>	<item 5>	<item
<item 0>	9117	0	0	0	0	0	0
<item 1>	0	107011	0	0	0	0	0

Figure A.1: Map Comparison Kit example of a Kappa comparison

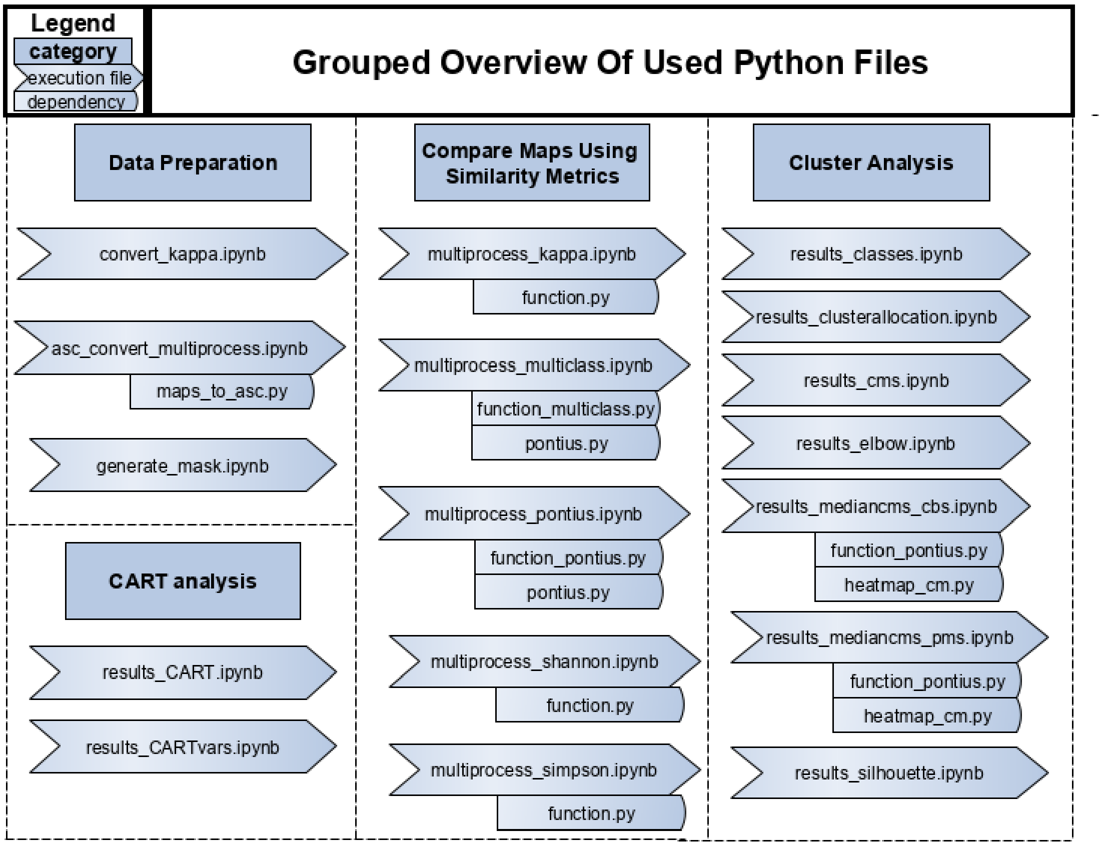


Figure A.2: Overview of used Python files

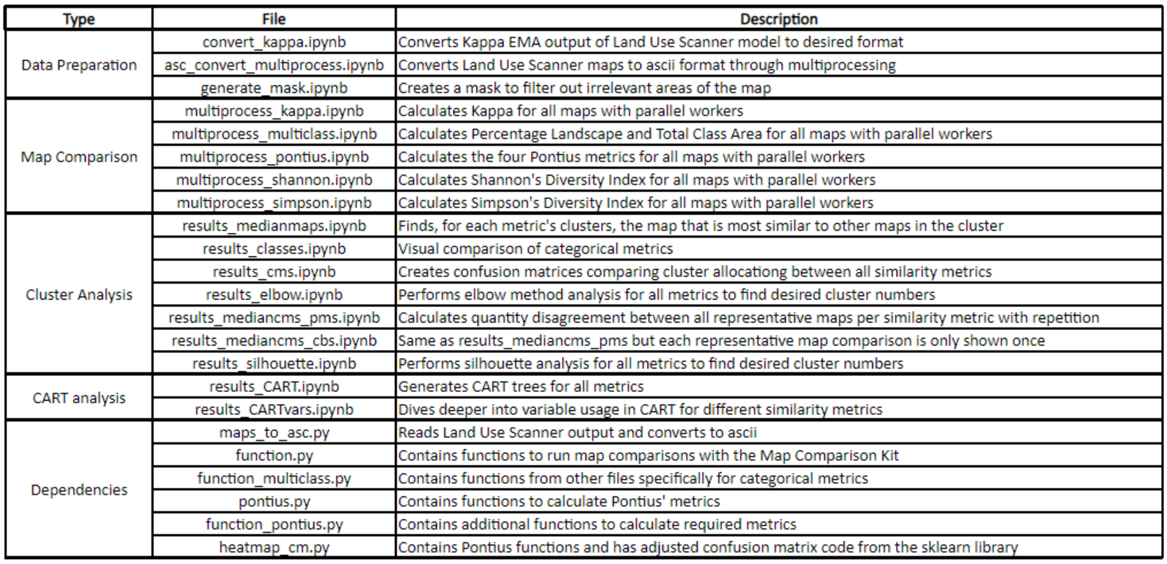
A.2 Python

Python played a large role in this research for processing the data and the results. All Python files used are available at `https://github.com/OQuispel/SD_similaritymetrics`. Setting up the experiments and generating results for the 2,000 maps was largely done as part of another study (Cox, 2020) and the relevant code can be found on the associated github, `https://github.com/margrietcox/LUS-scenario-discovery`.

A total of 23 Python files were used in this research. Their general purpose is presented and structured in Figure A.2. Table A.1 repeats this overview in a different format, while also providing a description of the purpose of the file. The files highlighted as dependencies are required files created specifically for the code in which they are used. The files pontius.py and function_pontius.py were retrieved from `https://github.com/verma-priyanka/pontiPy` and adjusted to fit the purpose of the research. Some adjustments were also made to the creation of confusion matrices from the sklearn library. A new version of this is added in heatmap_cm.py and in the function custom_cm.py, which greatly speeds up the generation of confusion matrices. Some other notable libraries required for some files are:

- Yellowbrick - used for the Elbow Method `https://www.scikit-yb.org/en/latest/`
- EMA workbench - used to generate results with LUS and also for CART analysis `https://emaworkbench.readthedocs.io/en/latest/`
- SciPy - used for clustering `https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html`

Table A.1: Overview of all Python files with description



Type	File	Description
Data Preparation	convert_kappa.ipynb	Converts Kappa EMA output of Land Use Scanner model to desired format
	asc_convert_multiprocess.ipynb	Converts Land Use Scanner maps to ascii format through multiprocessing
	generate_mask.ipynb	Creates a mask to filter out irrelevant areas of the map
Map Comparison	multiprocess_kappa.ipynb	Calculates Kappa for all maps with parallel workers
	multiprocess_multiclass.ipynb	Calculates Percentage Landscape and Total Class Area for all maps with parallel workers
	multiprocess_pontius.ipynb	Calculates the four Pontius metrics for all maps with parallel workers
	multiprocess_shannon.ipynb	Calculates Shannon's Diversity Index for all maps with parallel workers
	multiprocess_simpson.ipynb	Calculates Simpson's Diversity Index for all maps with parallel workers
Cluster Analysis	results_medianmaps.ipynb	Finds, for each metric's clusters, the map that is most similar to other maps in the cluster
	results_classes.ipynb	Visual comparison of categorical metrics
	results_cms.ipynb	Creates confusion matrices comparing cluster allocationg between all similarity metrics
	results_elbow.ipynb	Performs elbow method analysis for all metrics to find desired cluster numbers
	results_mediancms_pms.ipynb	Calculates quantity disagreement between all representative maps per similarity metric with repetition
	results_mediancms_cbs.ipynb	Same as results_mediancms_pms but each representative map comparison is only shown once
	results_silhouette.ipynb	Performs silhouette analysis for all metrics to find desired cluster numbers
CART analysis	results_CART.ipynb	Generates CART trees for all metrics
	results_CARTvars.ipynb	Dives deeper into variable usage in CART for different similarity metrics
Dependencies	maps_to_asc.py	Reads Land Use Scanner output and converts to ascii
	function.py	Contains functions to run map comparisons with the Map Comparison Kit
	function_multiclass.py	Contains functions from other files specifically for categorical metrics
	pontius.py	Contains functions to calculate Pontius' metrics
	function_pontius.py	Contains additional functions to calculate required metrics
	heatmap_cm.py	Contains Pontius functions and has adjusted confusion matrix code from the sklearn library

Appendix B

Large Scale Comparison of Maps

In this research two approaches were used for comparing large numbers of maps using similarity metrics. The first is a Python based implementation that reads model output and calculates the metrics directly. The second uses Python to run the Map Comparison Kit through command prompt. Both will be discussed. However, first the generation of the Land Use Scanner output will be discussed.

Note that all referenced Python files are available at `https://github.com/OQuispel/SD_similaritymetrics` unless stated otherwise.

B.1 Generating Land Use Scanner Output

The 2,000 maps that were used in this research for testing a variety of similarity metrics were obtained by running the Land Use Scanner over a wide range of the input space using the EMA workbench (J. H. Kwakkel, 2017). While some steps required for generating these results were performed in tandem with another student, most of the work was in fact performed as part of their study (Cox, 2020).

The input variables of the Land Use Scanner were set up so that 25 different binary variables can be enabled or disabled to impact model outcomes. These variables largely represent policies that impact how certain areas of land can, or cannot be used. Using these 25 variables, a subset of combinations was selected to generate a total of 2,000 different scenarios that the Land Use Scanner then processes en masse through application of the EMA workbench.

The code used for these steps can be found on the github of the research by Cox (2020):
`https://github.com/margrietcox/LUS-scenario-discovery`

B.2 Python Centered Implementation

A total of seven metrics are calculated directly through Python. These are:

1. Percentage Landscape
2. Total Class Area
3. Total Difference
4. Total Allocation Difference

5. Total Quantity Difference
6. Quantity Difference
7. Overall Accuracy

Percentage Landscape and Total Class Area are calculated in *multiprocess_multicass.ipynb*, while the remaining metrics are calculated in *multiprocess_pontius.ipynb*. The code used in the latter file was obtained from `https://github.com/verma-priyanka/pontiPy` and adjusted to suit the needs of this research. These notebooks follow the same approach for calculation of results, applied to different metrics. A function from a separate file is called using parallel workers. Each worker processes the calculation for a specific map pair, this allows for multiple comparisons to be run at the same time, limited by computational power available. As output is generated, these are stored in separate lists, one for each metric that the function calculates. When all possible comparisons are made, another function is called for each metric that processes the output previously stored in lists to a data frame, and exports it to disk as a csv file. This csv file is then used as the basis for cluster analysis.

B.3 Map Comparison Kit Implementation

Four metrics were calculated using the Map Comparison Kit. These are:

1. Kappa
2. Overall Accuracy
3. Shannon's Diversity Index
4. Simpson's Diversity Index

Note that most metrics could have easily been calculated with either the pure Python implementation or the Map Comparison kit, aslong as they are available in the Map Comparison Kit.

Kappa was calculated using *multiprocess_kappa.ipynb*, overall accuracy using *multiprocess_oa.ipyng*, Shannon's Index in *multiprocess_shannon.ipynb* and Simpson's Index in *multiprocess_simpson.ipynb*. The core approach is the same with the Python centered implementation, where each notebook calls on a function using parallel workers that each perform a single map pair comparison. The difference is that more steps are required than simply loading in the data. These can be summarized as follows:

- A number of different files are required by the Map Comparison Kit and must be made available:
  - a csl file must be generated, which describes what comparisons should be run
  - a log file must be available, which describes the maps that should be loaded into the program
  - files providing information on the data that should be ignored by the program needs to be provided (mask file)
  - the folder 'Legends' must be made accessible by the Map Comparison Kit
- Command prompt must be called through Python to run the Map Comparison Kit using very strict formatting
- After running a comparison the Map Comparison Kit generates an output statistic file. This file must be read so that the required statistic can be extracted and stored in Python.
- Following the previous step, a list is created as with the Python centered implementation, and this can be exported as a .csv in a similar manner.

B.4 Comparison of Calculation Methods

Both approaches have their merits and downsides. The main advantage of the Map Comparison Kit approach is that, when you get it running, you are working with a tested tool. A simple comparison of running the Map Comparison Kit through Python and doing it manually in the software itself will indicate whether it is working, and its previous use provides confidence that the results are correct. Most Python implementations are generally plugins for GIS software and hard to easily convert to a standalone source. Writing the code personally is not too challenging for most metrics, at least those analyzed in this research, however another program is likely still desirable to validate the results.

The Pontius metrics (Pontius & Santacruz, 2014) were conveniently available in a Python library, however some errors in the code were found that lead to wrong outcomes. Additionally, testing this library with another source is challenging as not all similarity metrics can easily be found elsewhere. For example, the Pontius metrics are not available in the Map Comparison Kit.

Beyond the validation aspect, the Map Comparison Kit is quite cumbersome to get up and running through command prompt. Not only does it require access to numerous files, minor errors can stop the whole process from working with little indication as to where the problem lies. However, once the infrastructure is there it is no different from running it in Python. A different matter is then whether calling on a separate program for a large number of calculations is desirable from a computational standpoint. While not tested in this research, it is likely that for large scale comparison of maps a separate efficient coding implementation is more efficient than calling on a separate program. Be this is in Python, or another programming language.

Appendix C

Finding Cluster Numbers

This appendix describes the process of clustering maps for the kappa process, going in depth into how the number of clusters were decided upon. Any code files referred to can be found at `https://github.com/OQuispel/SD_similaritymetrics`. Finding the cluster number results for other metrics can also be found in results folder. They are not all covered here in the same manner as kappa as the process is relatively subjective and the approach remains the same.

C.1 Clustering Process

In this research the hierarchical agglomerative clustering algorithm is used for all clustering purposes. The clustering process uses the csv files created in Appendix B. These files contain the distance values between each of the 2,000 maps. A dummy variable is then created for each similarity metric that clusters the maps on an initial number of clusters. This is then used in the elbow method and silhouette score to decide on a definitive cluster number. When this is found the clustering is rerun on said cluster number. This final variable can then be used for cluster analysis and CART analysis. The following sections will run through the described clustering process for the kappa metric to provide a better understanding on how the number of clusters were selected.

The code used for both the elbow-method and the silhouette score can be found in the files *results_elbow.ipynb* and *results_silhouette.ipynb* respectively.

C.2 Elbow Method

The first step undertaken in finding the desired number of clusters was to apply the Elbow method. The Elbow method recommends the number of clusters to use by plotting the explained variance against the number of clusters for a given data-set. The cut-off point where the diminishing returns of increased variance explanation do not outweigh using additional clusters is the 'elbow' and considered the recommended number of clusters in this method.

The visualization of the elbow method for the kappa metric clustering is presented in Figure C.1. The dotted line shows the recommended number of clusters according to the method. It is apparent that the preceding line for each cluster number before seven is steeper than its predecessor, with the exclusion of three clusters. Anymore than seven clusters leads to a less steep line., although an increase is seen after nine clusters. What the steepness indicates is the degree to which an additional cluster can explain away variance in the data. More clusters can always explain the data easier, however this can lead to over-fitting. While the recommended number of clusters seems plausible, the silhouette score was also calculated and visualized to confirm this choice.

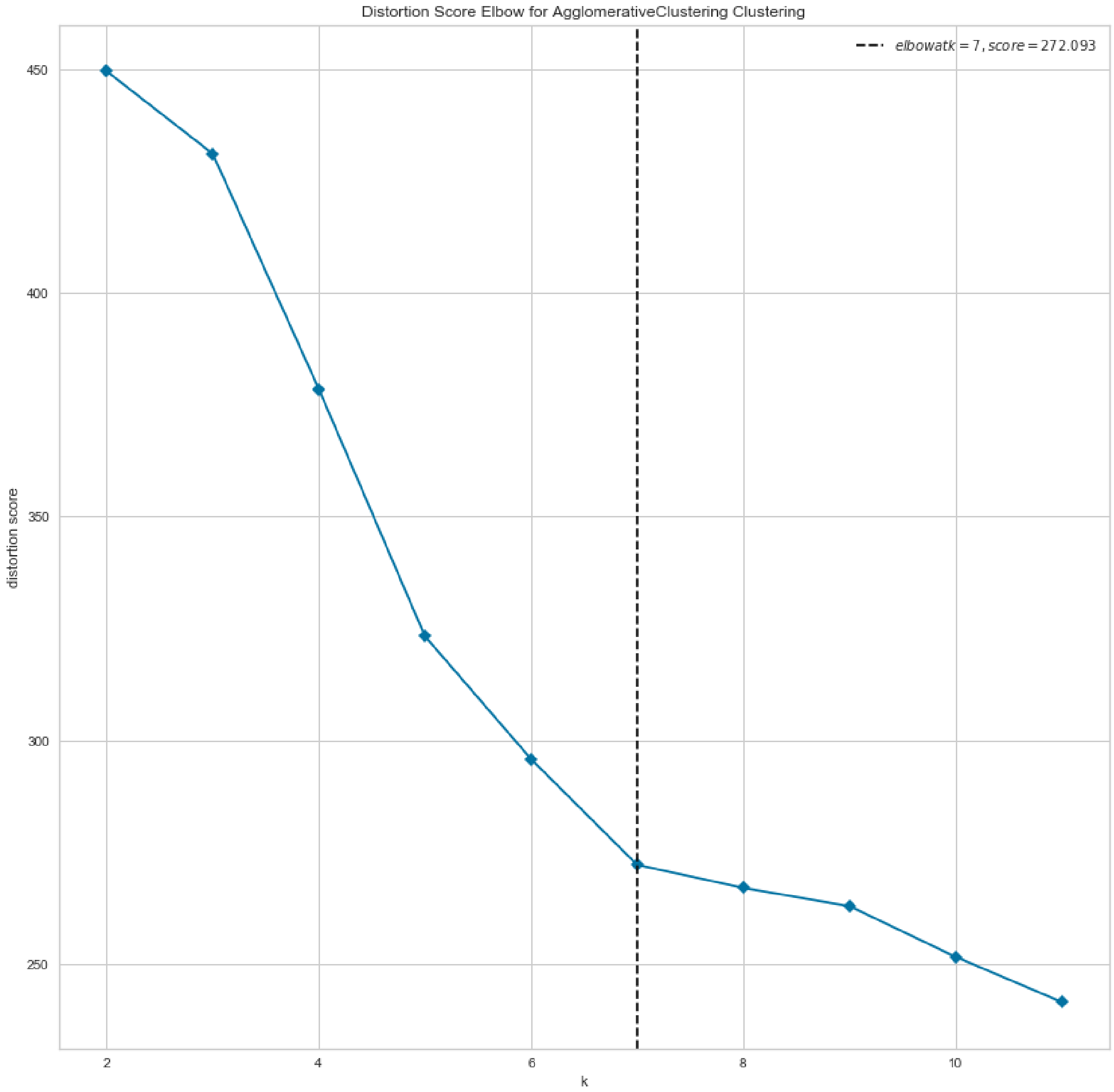


Figure C.1: Elbow Method results for the kappa metric

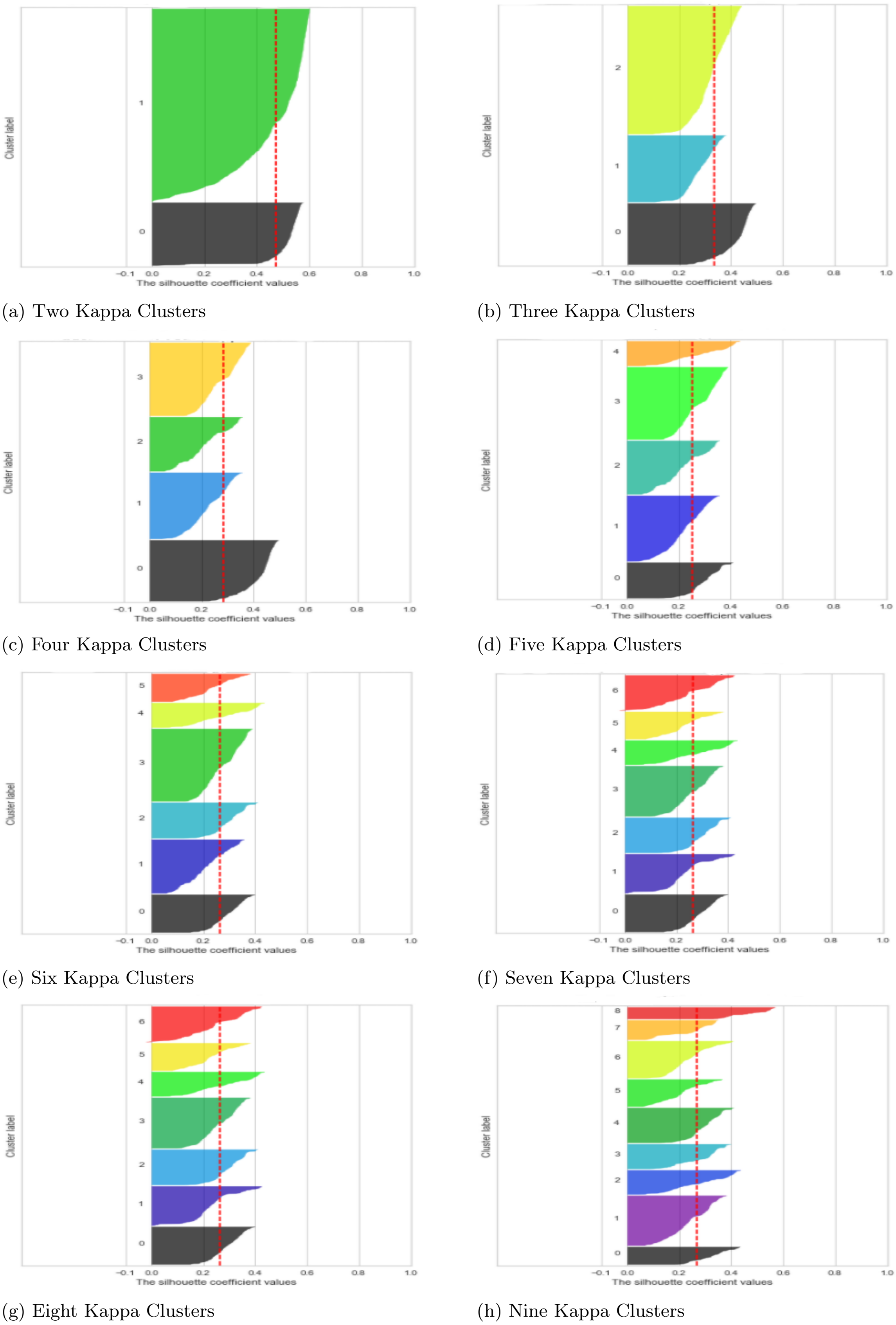
C.3 Silhouette Score

The Silhouette Score is a measure of how well each member of a cluster belongs to that cluster and other clusters. Similar to the elbow method, it provides a useful visual aid to represent this information. Figure C.2 presents eight different silhouette score graphs corresponding to cluster numbers two through nine for the kappa metric.

Each colored blob represents a cluster containing maps, in this case. The x-axis contains the silhouette score, where the higher the value the more a map can be considered to belong to its selected cluster over another cluster. It is apparent that the blobs each have unique shapes. This is because the silhouette score is computed for each individual entry, and not the cluster as a whole. In Figure C.2a this means that the black cluster contains maps with a similar silhouette score, while the green cluster, which is also larger as it is visualized wider, contains a relatively large number of maps that do not really belong in that cluster according to the silhouette score. As shown in C.2b, adding an additional cluster does not change the final cluster much, while the green cluster from Figure C.2a is seemingly partitioned into two clusters. This information is why the silhouette score is another useful tool for deciding on cluster numbers.

A clear trend is apparent where higher silhouette scores are achieved when there are fewer clusters. However, this comes at the cost of elements of each cluster not truly belonging to said cluster. As such a compromise must be made between in-cluster similarity (i.e. a more homogeneous silhouette score across each cluster) and the degree to which each map belongs to the cluster that it is assigned to, as judged by higher silhouette score on average.

While the elbow method provided a more stand-out result to choose as the optimal cluster number, the silhouette score provides more information for the user to process. Combining the two, the elbow method recommendation of seven clusters is followed as the silhouette score analysis shows that this value is an acceptable compromise between in-cluster similarity without over-fitting the data.



(a) Two Kappa Clusters

(b) Three Kappa Clusters

(c) Four Kappa Clusters

(d) Five Kappa Clusters

(e) Six Kappa Clusters

(f) Seven Kappa Clusters

(g) Eight Kappa Clusters

(h) Nine Kappa Clusters

Figure C.2: Overview of all silhouette score visualizations for kappa

Appendix D

Discovery and Analysis of Representative Maps

Chapter 4 presented the comparison of representative maps. This appendix will explain how these maps were found and compared. While chapter 4 provided the combinations of representative maps for the sake of brevity, this appendix will contain the full permutations. This provides the same information, but it was found that showing all respective map pairs with repetition makes the information easier to interpret.

All files referred to in this appendix can be found on the respective github:
`https://github.com/OQuispel/SD_similaritymetrics`

D.1 Representative Maps

The concept of representative maps was taken from another study that pioneered the use of scenario discovery in a land-use change modelling context by comparing scenario discovery results to the traditional story and simulation approach (Cox, 2020). The representative map for each cluster is defined as that map of which the sum of its difference with all the maps of its cluster is the lowest. The representative maps are first calculated in the file *result_medianmaps.ipynb*. This is achieved through the following steps:

1. Load the confusion matrix containing distance values for the similarity metric of interest
2. Filter the confusion matrix to only contain those maps that are part of the cluster for which the representative map must be found
3. For each row in the matrix, sum the distance values across all columns
4. The row with the lowest sum is the representative map for that cluster

Repeating this process for all clusters and metrics, provides early insight into how the different clusters vary, both for clusters of a metric as between metrics in general. An overview of the representative maps found using this process was first presented in Chapter 5, and can also be found in Table D.1.

Table D.1: Overview of representative maps for each similarity metric

Metric	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
Kappa	Map 359	Map 456	Map 541	Map 695	Map 1148	Map 1760	Map 1911
Overall Accuracy	Map 10	Map 359	Map 456	Map 541	Map 695	Map 696	Map 1911
Total Difference	Map 359	Map 456	Map 541	Map 695	Map 1148	Map 1760	Map 1911
Total Allocation Difference	Map 263	Map 453	Map 678	Map 1760	Map 1852	Map 1918	-
Simpson's Diversity Index	Map 65	Map 857	Map 1448	Map 1784	Map 1830	-	-
Total Class Area [Corn]	Map 246	Map 296	Map 325	Map 825	Map 1612	-	-
PLAND [Corn]	Map 246	Map 325	Map 564	Map 704	Map 843	-	-
Total Quantity Difference	Map 1057	Map 1472	Map 1833	Map 1978	-	-	-
Shannon's Diversity Index	Map 61	Map 1248	Map 1454	Map 1457	-	-	-
Total Class Area [Residential]	Map 217	Map 332	Map 604	-	-	-	-
PLAND [Residential]	Map 264	Map 1422	Map 1877	-	-	-	-
Total Class Area [Nature]	Map 0	Map 8	Map 35	-	-	-	-
PLAND [Nature]	Map 0	Map 8	Map 35	-	-	-	-
Quantity Difference [Nature]	Map 0	Map 1	Map 8	-	-	-	-

D.2 Comparison of Representative Maps

Multiple approaches can be used for comparing the representative maps. As comparing the maps visually did not prove fruitful, a numerical approach was investigated. After some trial and error it was found that creating a heatmap based on a similarity metric value that compares each representative map pair proved useful.

The quantity difference metric was selected for this purpose (Pontius Jr & Millones, 2011). The reason that quantity difference was selected is that it is simple to calculate and interpret, while providing useful information. This metric, when normalized, provides the percentage of the total study area that map one and map two of the comparison assign differently from one another for a given land-use category. This metric thus highlights a focus with regards to the quantity of each land-use category. What it does not provide is any information on whether the allocation of these categories is in the same location.

The process to achieve the desired comparisons is as follows:

1. Create the list of representative maps to be compared
2. Calculate and store the quantity difference value for each pair of maps for each land-use category in the model
3. Convert the data into a dataframe where the rows show the map pairs, the columns represent the land-use categories, and the values are the result of the quantity difference metric
4. Apply a heatmap to the dataframe to highlight differences of note

Following this procedure provides a visual tool to compare the representative maps for a given similarity metric, and find whether different narratives exist between clusters. These figures were first presented in chapter 4 where each map pair was shown without repetition. Following, instead the permutations of all map pairs are shown as this better highlights the differences between maps at the cost of more space. Note that the interpretation, while easier to obtain with the permutations, does not change when looking only at the combinations.

The files containing the creation of the permutation and combination figures can be found in *results_mediancms_pms.ipynb* and *results_mediancms_cbs.ipynb* respectively. An excel containing all permutations for all land-use categories is also available in *heatmap_qd_permutations.xlsx*. The representative map comparisons as presented in by land-use category in Chapter 4 are created in *results_mediancms_cat.ipynb*.

Appendix E

Overview of Cluster Allocation Confusion Matrices

This appendix discusses the cluster allocation confusion matrices that were first presented in chapter 5. The purpose of these matrices is to visualize the difference or agreement in cluster allocation when using different similarity metrics. As such, each confusion matrix is the result of comparing the cluster allocation results for two metrics. First a short introduction will be provided on the creation of the confusion matrices, followed by a presentation of all confusion matrices.

E.1 Creation of Cluster Allocation Confusion Matrices

The confusion matrices are created by passing the clustering labels for each metric to the sklearn confusion matrix function. The confusion matrix is then converted to a dataframe to make it easier to process. The seaborn library is used to apply a heatmap to the dataframe. This process is automated for all metric pair combinations by passing a list to a function which can process the list in parallel with a set number of workers.

This process is applied in *results_cms.ipynb*.

E.2 Presentation of all Confusion Matrices

Due to the large number of confusion matrices they are instead available on the github in the *cms* folder. Interpretation is left to the reader, as the noteworthy examples were presented earlier in chapter 5.

Appendix F

CART Analysis

Classification and Regression Tree (CART) analysis was used in this research as the rule induction algorithm to retrace outcomes to specific input parameters. This was presented in chapter 5 for three outcomes considered interesting. Additional analysis was also presented regarding the usage of input variables between similarity metrics, and CART tree attributes were also described. This appendix will show all CART trees, including those not presented in chapter 5. As the CART trees in this research are mostly of interest relative to another tree, interpretation is left to the reader.

CART was applied using the EMA workbench (J. H. Kwakkel, 2017). The used code was first applied in another study that compared story and simulation to scenario discovery in land-use change modelling (Cox, 2020). The code used in this research to generate the CART trees can be found in the file `results_CART.ipynb`. Some basic, but useful, analysis performed to provide insight in the model variables that the CART analysis highlighted can be found in `results_CARTvars.ipynb`. This file turns an excel file that was used to manually track variable usage by CART trees into a visually more appealing output, as presented in

The raw CART trees generated in Python are all available on the github `https://github.com/OQuispel/SD_similaritymetrics` in the *cart_trees* folder.