Replication of agent-based models in archaeology

A case study using Brughmans and Poblome’s MERCURY model

Hilde Kanters

Replication of agent-based models in archaeology: a case study using Brughmans and Poblome's MERCURY model

Hilde Kanters, s1272543

Master thesis

Dr. K. Lambers

Digital Archaeology

University of Leiden, Faculty of Archaeology

Leiden, 24-10-2018, final version

Table of Contents

1 Introduction 6

1.1 Replication and agent-based modelling 6

1.2 Methodology: the background of agent-based modelling 8

1.3 Research questions 11

2 MERCURY and its results 16

2.1 The archaeological context of MERCURY 16

2.2 A detailed explanation of MERCURY 18

2.3 Brughmans and Poblome's results and conclusions 26

3 The replication process and its results 33

3.1 Version 1 35

3.2 Version 2 38

3.3 Version 3 41

3.4 Version 4 43

3.5 Version 5 48

3.6 Version 6 54

3.7 Version 7 57

3.8 Version 8 59

3.9 Statistical comparison 61

3.10 Discussion 64

4 Critiques of MERCURY 67

4.1 Existing critiques of MERCURY 67

4.2 New critiques of Brughmans and Poblome's research 69

4.3 Additional experiments and their results 75

5 Discussion 78

6 Conclusions 83

Abstract 87

Internet Pages 88

Bibliography 89

List of Figures 98

List of Tables 101

List of Appendices 103

Acknowledgements

I would like to thank Tom Brughmans and Jeroen Poblome for publishing the research on which this thesis is based. In addition, I would like to thank Iza Romanowska, Fulco Scherjon and Karsten Lambers for introducing me to agent-based modelling, and computer programming in general, through their course on simulation in archaeology. Karsten Lambers also supported me as my thesis supervisor and Fulco Scherjon provided me with help during the coding process of this thesis, for which I am grateful. Lastly, I would like to thank Jonathan Ozik for maintaining the Repast mailing list of which I directly and indirectly made use.

1 Introduction

1.1 Replication and agent-based modelling

Agent-based modelling, or ABM for short, is a tool used to study complex systems through the simulation of agents interacting with each other and their environment. ABM has been used as an instrument for scientific research in various fields, such as computer science, management, various social sciences, economics, geography and, of course, archaeology (Macal 2016, 146-147). Although this methodology is sometimes described as being new, it has been used for more than 40 years (Lake 2014a, 6). The simulation study of Palaeolithic social systems by Wobst (1974), can be called one of the first agent-based modelling applications in archaeology. Since the turning of the millennium, the amount of published simulation studies has expanded dramatically, and the method has, arguably, obtained some form of 'maturity' (Lake 2014b, 277-278).

However, the surge in popularity of agent-based modelling brings with it a methodological problem. Even though ABM is becoming more and more common, replication studies of archaeological simulations are virtually non-existent. Replication can be defined as the reproduction of a published experiment, generally by scientists independent of those who performed the original study, based on the published details of the research. If the reproduced experiments are determined to be similar enough to the original, generally through the use of statistical tests, it can be called a successful replication. Replication studies are an important factor of scientific research as they allow us to check whether the published descriptions of experiments are accurate and the results are not reliant on local conditions (Wilensky and Rand 2007). Wilensky and Rand (2007) argue that replication is even more important in computer simulation than it is in physical experimentation, as it cannot only show that an experiment is not a one-time event, but it can also bring more confidence to the model verification, whether the implemented agent-based model reflects the conceptual model on which it is based, and validation, whether the implemented model fits the real-world processes it tries to simulate. The lack of replication studies leads some to believe we might soon face a time in which the validity of existing models will be doubted (Romanowska 2015b, 186). In Romanowska's (2015b, 186) own words: "*[I]t is only by replicating simulation studies, constructing libraries of tested models and reusing them, and continuously challenging the models with new data and new hypotheses that a high level of certainty can be obtained. Given the current hype around simulation models in archaeology, and the relative scarcity of replication studies, we may expect a turbulent but necessary period of questioning the existing models to follow soon.*" To paint a picture of this scarcity: the only replication study of an archaeological agent-based model that I was able to find was one of the 'Artificial Anasazi' model (Janssen 2009), arguably the most well known archaeological ABM. It should be noted that this issue is not at all unique to ABM in archaeology, as even outside of the field of archaeology, the vast majority of agent-based models remain unreplicated (Wilensky and Rand 2007). The replication of computational archaeological research outside of ABM is also scarce, as pointed out in a study by Marwick (2017), wherein he aims to address this issue by creating a standardised way of publishing research in order to facilitate replication.

Simulation studies have been criticised by parts of the archaeological community. Such criticisms include simulations being deterministic, reductionist and being incapable of incorporating the subjectivity of human behaviour (Lock 2003, 148-149), as well as being intransparant 'black boxes', which hide information (Huggett 2004, 83-84). The field of simulation has also been criticised for fetishizing new and innovative technologies and for being predominantly male (Huggett 2004, 82-88). Strengthening ABM methodology through replication, could help to convince critics of the validity of simulation in archaeology.

The small amount of ABM replication studies from outside the field of archaeology have shown why they are so important. For example, the study by Will (2009) shows the interesting and important results replication can yield. Without going into too much detail, the model that was replicated concerns social mobility and market formation and was used to compare the individualist USA with collectivist Japan. Will (2009) found that one assumption, which was made explicit in the code of the original model, was not justified in the corresponding papers. When this assumption was left out of the model, the results differed greatly from the original. However, the original creator's of the model responded to this study by recalibrating one input variable in the replicated model, which then, surprisingly, resulted in a better fit with their hypothesis than the original model did (Macy and Sato 2010). Other replication studies have shown shortcomings in the original model (Edmonds and Hales 2002; Miodownik *et al*. 2010) or reinforced the importance of documentation (Donkin *et al*. 2017). Some replication studies are almost directly 'successful', and do not significantly contradict the original model (Axtell *et al*. 1996; Janssen 2009). However, these studies are still important to publish as they allow us to put more trust in the original models.

It is clear that the lack of replication studies is a significant problem that should not be ignored. Therefore, I aim to address this problem in my thesis. Of course it will be impossible to test a large amount of models in the timespan available to me. However, it will be possible to show the procedures that have to be followed during model replication and highlight the importance of model documentation, in addition to replicating and thoroughly examining a single model. As my personal experience with agent-based modelling prior to writing this thesis was limited, having only followed one short university course on the subject, it will hopefully also show the feasibility of learning to replicate simulation studies to fellow archaeologists.

1.2 Methodology: the background of agent-based modelling

Of course, agent-based modelling will be the main method used in this thesis. Therefore I will now briefly describe the aspects of this method.

An agent-based model is in essence a computer model in which agents interact with each other, and optionally the environment in which they exist, based on predetermined rules, resulting in a complex system. An 'agent' in the context of agent-based modelling can be described as an object that, firstly, can act autonomously based on a range of preset rules, secondly, has certain traits or features that influence its actions, and, thirdly, has interactions with other agents. An agent may have other characteristics, such as existing in an interactable environment, having specific goals which govern its behaviour, the ability to learn and adapt over time and possessing certain resources, such as money or energy (Macal and North 2009, 87-88). A complex system can generally be defined as a system in which individuals, agents in the case of agent-based modelling, interact with one another to produce results that cannot be simply deduced from their actions. An example of a complex system is the Darwinian idea that, through interaction, simple organisms evolve into more complex and specialised one's (Heath and Hill 2010, 163).

Agent-based modelling emerged from the field of complex adaptive systems (Heath and Hill 2010). This field of study covers the way in which the interaction between autonomous agents results in complex systems, with the primary axiom that these systems emerge from the ground-up (Macal and North 2009, 88-89). There are seven characteristics of complex adaptive systems that have been identified by Holland (1995), which were fundamental in the development of agent-based modelling as a field (Heath and Hill 2010, 167-168). These are:

- Aggregation: the ability for subgroups to form
- Tagging: the capability of subgroups, agents in the case of ABM, to be recognised
- Building blocks: the re-use of subgroups to form different patterns
- Non-linearity: the notion that the results of a complex adaptive system are not the same as the sum of its components
- Flow: the transference of information between agents
- Internal models: the rules that govern the behaviour of agents
- Diversity: even under the same external conditions, different agents will not behave in a uniform way

Another important aspect of complex adaptive systems and agent-based modelling is emergence. Emergence can be described as the manifestation of new, macroscopic, features from the lower-level interaction between agents. The precise form of emergent properties are not able to be deduced from the interaction between agents from which they arise (Epstein and Axtell 1996; Goldstein 1999, 50). A common example of emergence is the shape of bird flocks. Individual birds adhere to certain rules, such as avoiding collision and matching flight speed with other birds in their immediate surroundings, which determines the shape of a flock as a whole (Hermellin and Michel 2017).

Agent-based modelling is a method that can be used to study the mechanisms of complex systems described above. By programming the actions of agents, resulting in the emergence of a complex system, the rules and variables which allow for this emergence can be studied.

A classic example of the use of ABM in archaeology is the Artifical Anasazi model, which was used to study settlement patterns of the Anasazi in Long House Valley, Arizona (Axtell *et al*. 2002; Dean *et al*. 2000; Gumerman *et al*. 2003). Through simple rules, the agents in this model, which represent households, interact with one another and the environment and choose settlement locations. Variables relating to demographic numbers, social relations and interaction and environmental conditions are included in this model (Gumerman *et al*. 2003, 436). The rules of interaction result in the emergence of settlement patterns that are comparable with archaeological data. This model was used to show that the decline and abandonment of Long House Valley cannot be solely attributed to environmental change, but also to social pull factors (Gumerman *et al*. 2003, 442-443).

Other applications of agent-based modelling in archaeology include the study of: Pleistocene human dispersal (Callegari *et al*. 2013; Cuthbert *et al*. 2017; Romanowska 2015a; Scherjon 2012), farming and pre-industrial economic production (Angourakis *et al*. 2014; Barton *et al*. 2010; Cockburn *et al*. 2013), historical societal collapse (Arikan 2017), social interaction and change in hunter-gatherer and nomadic groups (Barceló *et al*. 2014; Briz i Godino *et al*. 2014; Clark and Crabtree 2015), the emergence of social hierarchy (Crabtree *et al*. 2017; Rouse and Weeks 2011), Palaeolithic lithic procurement (Brantingham 2003), mobility in hunter-gatherers (Santos *et al*. 2015), (pre-)historical warfare (Cioffi-Revilla *et al*. 2015; Turchin *et al*. 2013), trade (Brughmans and Poblome 2016a; Crabtree 2016; Ewert and Sunder 2018), prehistoric seafaring (Davies and Bickler 2015), archaeological deposit formation (Davies *et al*. 2015), the division of labour in Iron Age salt mines (Kowarik *et al*. 2012) and archaeological field surveys (Rubio-Campillo *et al*. 2012).

1.3 Research questions

The model that I have chosen to replicate is the MERCURY model by Tom Brughmans and Jeroen Poblome (2016a; 2016b). MERCURY stands for Market Economy and Roman Ceramics Redistribution. As the name suggests, this model was created to explore the complex aspects of the economy of the Roman Empire. Although the model could be used in a broader context, Brughmans and Poblome (2016b) limit their research to the Eastern Mediterranean from 25 BCE to 75 CE. This period was focused on because the archaeological tableware data from this area, which was used to compare the simulated data to, shows a particular pattern of interest. This will be explained in more detail in chapter two.

The MERCURY model was chosen to be replicated because it is a exemplary case of ABM in archaeology as it includes two important features: hypothesis testing and comparison with archaeological data. The two hypotheses that were tested using this model are by Bang (2008) and Temin (2012) and they concern the workings of the Roman economy. According to Bang's bazaar hypothesis, the integration of markets was weak and access to information concerning supply and demand was limited, resulting in a more fragmented economy. Bang's main methodology is comparative history; an elaborate comparison between the Roman Empire and the Mughal Empire is made. Although Bang (2008) does not clearly state that his book is limited to a certain period within the history of the Roman Empire, he mostly discusses the early Roman Empire. Temin (2012) too focuses on the early Roman Empire. In contrast to Bang, Temin's view of the Roman economy is one in which commercial information is able to flow more freely throughout different communities, with the existence of one large market as a result. Both authors draw on a plethora of historical and archaeological sources from across the whole Roman Empire. In their ABM study, Brughmans and Poblome (2016a, 395-397) use different parameter settings, representing the two hypotheses, and compare the distribution pattern of their output data to a distribution pattern found in an existing database of over 33000 sherds of Eastern Roman tableware.

In order for a replication to produce significant results, it should differ in certain ways from the original in terms of implementation. Wilensky and Rand (2007) identify six ways in which a replication can differ from its original: the time at which a simulation is performed, the hardware that the simulation is ran on, the language the model is written in, the toolkit that was used when writing the code, the specifics of the algorithms that are used and in which order they operate, and the authors of the models. The time, hardware and the author of the model will necessarily differ from the original model, in this case. In addition, the decision was made to use a different toolkit and coding language to program the model. The algorithm was not specifically chosen as a way in which the replicated model will differ from the original, but it is possible that it will also differ in this aspect, as the ODD, will be followed as the main guide when writing the replication instead of the source code. The ODD, short for 'Overview, Design concepts and Details', is protocol designed to standardise descriptions of agent-based models and aid in replication attempts (2006; 2010). The ODD consists of: an *overview* of the model, including its purpose, variables and process overview; a section on *design concepts*, in which certain aspects of the model, such as the stochasticity, emergence and the ability of agents to learn and interact, can be explained; and a section on the *details* of the model's processes and its input data.

The authors of the original MERCURY model used the NetLogo toolkit (Brughmans and Poblome 2016b). NetLogo (ccl.northwestern.edu, b) is a programming language and toolkit which was specifically designed for agent-based modelling. For this replication, I have chosen Repast as a substitute. Repast (repast.github.io, a) is an agent-based modelling suite that can be divided into two distinct versions: Repast HPC and Repast Simphony. Repast HPC, which stands for High Performance Computing, uses the C++ language and is designed for complicated models running on large clusters of computers or supercomputers. Repast Simphony is a more accessible version that can use a combination of the languages Java, Groovy and ReLogo, which is a language specifically designed for agent-based modelling, comparable to NetLogo (Ozik *et al*. 2013, 1560-1561). For this replication, the Repast Simphony toolkit (version 2.4) was chosen, in combination with the Groovy and ReLogo programming languages. This decision to use these languages was primarily made for convenience, as ReLogo is similar in terms of syntax to NetLogo, with which I already have experience, and Groovy is described as a more accessible alternative to Java.

There are three categories of replication standards that can be met: numerical identity, distributional equivalence and relational alignment (Axtell *et al*. 1996, 135). Numerical identity is the exact equivalence of numerical output. Due to the stochastic nature of most agent-based models, this will be impossible to prove in almost all cases, as even the same model can produce slightly different numerical results using the same parameter settings. In stochastic models, the only way numerical identity could be achieved is to use the exact same software and use the exact same random number generator settings. The replicated and original model are said to be distributionally equivalent if the output is statistically indistinguishable from one other. Although Axtell *et al*. (1996) do not give mention specific statistical tests that could be used to test for distributional equivalence, the one's they use are Mann-Whitney U test and Kolmogorov-Smirnov tests. Other studies citing this replication standard, t-tests of various kinds are used (Donkin *et al*. 2017; Wilensky and Rand 2007). Relational alignment, the weakest replication standard, is achieved when the models' output data and input variables show the same relationship between them. Because of the stochastic elements in the MERCURY model, and because a different ABM toolkit is used, numerical identity, the strongest replication standard, can not be achieved. Therefore, in this study, distributional equivalence will be aimed for, because it is deemed to be a stronger standard than relational alignment (Axtell *et al*. 1996, 135).

The experiments presented in the supplement 1 of Brughmans and Poblome (2016b) will be replicated using the same input variable values. Ideally, the replicated model and the original model will be declared as matches if they pass statistical tests for equality of the mean or distribution, such as a paired samples t-tests for example. The specific test used, will depend on the properties of the output data, like the normality of the distribution. Using such statistical tests to test whether the replicated model and the original model 'match' is customary in replication studies (Axtell *et al*. 1996; Donkin *et al*. 2017; Edmonds and Hales 2002; Miodownik *et al*. 2010; Wilensky and Rand 2007, 146-149). Naturally, if the models do not initially match, the source of this mismatch will be sought. This involves a stepwise execution of checking the code manually and comparing it to the source code and performing subsequent statistical analyses when changes are made to the code.

Although the amount of data that was included in the papers by Brughmans and Poblome (2016a; 2016b) is much greater than that of other archaeological ABM studies I've looked at, it does include the data that is necessary to perform adequate statistical tests. Brughmans and Poblome (2016b, supplement 1) only reported the means of 100 simulation runs for each of the 35 experiments, but not the output data of each individual run. In an email exchange, Tom Brughmans kindly provided me the necessary output data to compare the tableware distributions simulated by MERCURY (appendix 1). Sadly, this data did not include network measures, as they were only performed for one experiment in every 100 of its kind. Therefore, network measure data of the replication cannot be compared to the original statistically, only the descriptive statistics of each experiment can be compared. However, since the network structure influences the tableware distribution, but not vice-versa, statistical tests of the tableware distribution will also say something about the networks. If this explanation is confusing, it will be clear after reading the next chapter on the intricacies of the MERCURY model and its results.

The research questions that I aim to answer in this replication study are:

- Can an independent replication of the MERCURY model match the results presented by Brughmans and Poblome (2016a) on a distributional level, as defined by Axtell *et al* (1996)?

- Can this replication be performed based solely on the description in the ODD, and if it cannot, what are the shortcomings of the ODD?

- If the models cannot be matched, what causes the differences between them?
- What consequences, if any, will this replication attempt have on the original study by Brughmans and Poblome (2016a; 2016b)?
- How does this replication of MERCURY compare to other replication studies?

Specific emphasis is given to replication using the ODD as a guide. The ODD protocol was designed by Grimm *et al.* (2006; 2010) as a standardised way of describing agent-based models. Emphasis is given to the ODD not only because one of its main aims is to assist in replication, but also because it should contain a detailed explanation of the model that does not rely on pre-existing knowledge of a specific programming language. Readers should be able to rely upon the ODD if the explanation of a model in the published paper is insufficient. The accuracy of the ODD can not be confirmed if only the source code is used as a guide in replication process.

2 MERCURY and its results

In this chapter I will explain the MERCURY model and the conclusions Brughmans and Poblome (2016a; 2016b) have drawn from its experiments. This is of course already done in the publications by the original authors, but for the sake of completeness I believe it to be necessary to include a description here as well. The functions, variables and constants of the model will have to be described to make the subsequent chapters on the replication of the model and critiques of it comprehensible. Unless otherwise noted, the details about the model are from the ODD found on the MERCURY page at CoMSES Net / OpenABM (www.comses.net, a).

2.1 The archaeological context of MERCURY

Brughmans and Poblome (2016a; 2016b) created their model to study the distribution patterns of *terra sigillata* tableware throughout the Eastern Mediterranean. By examining a dataset from the ICRATES project of over 19 700 sherds, described in Bes and Poblome (2008), Brughmans and Poblome (2016b, 395-397) observed a pattern in the distribution width and range of the tableware types Eastern Sigillata A, B, C and D. Due to the limitations of the dataset, critical quantitative analysis was not performed; only broad distribution patterns were assessed. Brughmans and Poblome found that between 25 BCE and 75 CE, Eastern Sigillata A dominated the assemblage. It had by far the widest distribution of the four types until 75 CE. From 100 to 150 CE, Eastern Sigillata D overtakes Eastern Sigillata A as the dominant tableware type, although the degree of its dominance is not as extreme as Eastern Sigillata A was before (fig. 1). Brughmans and Poblome (2016a) formulated the following research questions which they aimed to answer using their agent-based model: "*What hypothesised processes could give rise to this pattern? How does the* *availability of reliable commercial information to traders affect the distribution patterns of tableware?*" In this question, 'this pattern' refers to the dominance of one pottery type over the others, not to the shift in dominance from one type to another.

In order to address this question, Brughmans and Poblome (2016a; 2016b) used the MERCURY agent-based model to make explicit and compare two conceptual models that might explain the observed pattern: the 'Roman bazaar' model by Bang (2008) and the 'Roman market economy' model by Temin (2012). In short, Bang (2008, 4) describes his Bazaar model as follows: "*Compared to modern markets, the bazaar is distinguished by high uncertainty of information and relative unpredictability of supply and demand. This makes the prices of commodities in the bazaar fairly volatile. As a consequence, the integration of markets is often low and fragile; it is simply difficult for traders to obtain sufficiently reliable and stable information on which effectively to respond to developments in other markets. Considerable fragmentation prevails.*" In contrast, Temin's (2012, 4) view of the Roman economy involves large, empire-stretching markets: "*I argue that the economy of the early Roman Empire was primarily a market economy. The parts of this economy located far from each other were not tied together as tightly as markets often are today, but they still functioned as part of a comprehensive Mediterranean market.*" Another quote by Temin (2012, 17) shows that he believed there was a much freer flow of information throughout the market than Bang did: "*While the demand for Roman wheat might have risen, each Sicilian or Egyptian farmer would only have known what price—or tax rate—he faced. We have several surviving comments about the prevailing price of wheat, some in normal times and more in unusual ones. The presence of these prices indicates that both farmers and consumers knew what the price was. Since these prices typically were not for individual transactions, they also indicate the presence of anonymous exchanges. We have no way of knowing how widespread this information was, but the quotations suggest strongly that this was general information. It makes sense therefore to see farmers as facing a competitive market in which their output was too small to affect the price. They then made their choices on the basis of what they saw as a fixed market price, just as farmers do today.*"

These stark contrasts between the two conceptual models were also described

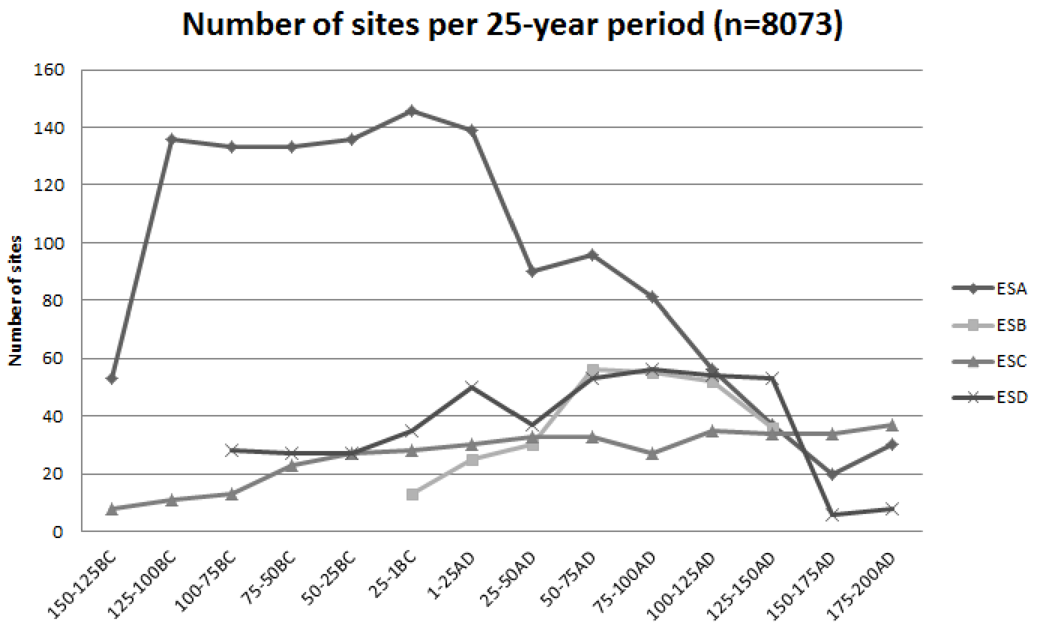

Figure 1: A graph showing the amount of sites each tableware type was found on, based on ICRATES data (Brughmans and Poblome 2016a).

by Brughmans and Poblome (2016a; 2016b). The differences were made explicit in the MERCURY model by changing input variables to reflect the two conceptual models. I will get back to this after describing the specifics of the MERCURY model. There are other differences between Bang (2008) and Temin's (2012) models, such as the importance of social relations and state influence, that were not incorporated into MERCURY (Brughmans and Poblome 2016a).

2.2 A detailed explanation of MERCURY

In essence, the MERCURY model represents trade networks of the Roman Empire. There are two types of agents that are essential to the MERCURY model: sites and traders. The traders take on an active role, as they exchange products between each other based on predetermined rules. Sites are passive; they store discarded and traded goods and a subset of them, the production sites, allow traders that are located there to 'produce' new tableware. There also exists a third entity: links. Links determine which traders are connected. Only linked traders can exchange information and trade products with each other. I would argue that links are not agents in this case, as they do not perform actions. They only provide the network structure which is used by traders. The model does not have a sense of scale. Links between traders are the only representation of space, but they do not represent geographical distance, nor is there a difference between the amount of space different links represent. Neither does each time step represent a certain amount of time, such as days or months (www.comses.net, a). At the end of a run, network measures and the amount of sites each tableware type is spread on serves as the output data.

Table one and two contain all independent and dependent variables used in the MERCURY model, the independent variables being the input values, which influence the creation of the network and the actions of agents, and the dependent variables being the values in which information is stored and which change throughout the simulation.

At the start of each simulation a number of sites are created equal to the *num-sites* value and a number of traders are created equal to the *num-traders* value. These values are 100 and 1000 respectively in every experiment performed by Brughmans and Poblome (2016a). These sites are visually aligned in the shape of a circle. Four of the 100 sites are chosen to be productions sites, i.e. their *production-site* variable and one of the *producer-X* (tab. 1) variables are set to 'true'. There is one production site for each of the four products: A, B, C and D. The production sites are equally spaced along the circle (Brughmans and Poblome 2016a). The distribution of the 1000 traders among the sites is dictated by the *equal-traders-production-site*, *traders-distribution* and *traders-production-site* independent variables. If *equal-traders-production-site* is 'true' an equal amount of traders, the number of which is determined by the *traders-production-site* variable, is moved to each production site first. Afterwards, all other traders are distributed over the remaining non-production sites. If *equal-traders-production-site* is 'false', production sites are treated the same as non-production sites for the purpose of trader distribution, except if *traders-production-site* variable is equal to '30,1,1,1'. In this case, 30 traders are distributed to production site A and the other production sites are assigned one trader per site (Brughmans and Poblome 2016a). The variable *traders-distribution*

Table 1: Independent variables (after Brughmans and Poblome 2016a, supplement 2).

Variable	Description	Tested Values
	Global Variables	
num-traders	The total number of traders to be distributed among all sites	1000
num-sites	The total number of sites	100
equal-traders-production-site	Determines whether the number of traders at production sites will be equal and determined by the variable *traders-production-site* or whether it will follow the same frequency distribution as all other sites determined by the variable *traders-distribution*	true, false
traders-distribution	Determines how the traders are distributed among the sites	exponential, uniform
traders-production-site	Determines the number of traders located at production sites if *equal-traders-production-site* is set to 'true'	1, 10, 20, 30
network-structure	Determines how the social network is created when initialising an experiment: a randomly created network, or the network structure hypothesised by Bang or Temin.	hypothesis, random
maximum-degree	The maximum number of connections any single trader can have	5
proportion-inter-site-links	The proportion of all pairs of traders that are connected in step two of the network creation procedure by inter-site links	0; 0,0001; 0,0006; 0,001; 0,002; 0,003
proportion-intra-site-links	The proportion of all pairs of traders that are considered in step three of the network creation procedure to become connected by intra-site links	0.0005
proportion-mutual-neighbors	The proportion of all pairs of traders with a mutual neighbour that are considered for becoming connected in step four of the network creation procedure by intra-site-links	2
	Site-specific variables	
production-site	Set to 'true' if the site is a production centre of one of the products	true, false
producer-A	Set to 'true' if the site is the production centre of product-A	true, false
producer-B	Set to 'true' if the site is the production centre of product-B	true, false
producer-C	Set to 'true' if the site is the production centre of product-C	true, false
producer-D	Set to 'true' if the site is the production centre of product-D	true, false
	Trader-specific variables	
max-demand	The maximum demand each trader aims to satisfy	1, 10, 20, 30
local-knowledge	The proportion of all link neighbours a trader receives commercial information from (supply and demand) in each turn	0,1; 0,5; 1

Table 2: Dependent variables (after Brughmans and Poblome 2016a, supplement 2).

Variable	Description
	Site-specific variables
volume-A	The number of items of product A deposited on the site as a result of a successful transaction
volume-B	The number of items of product B deposited on the site as a result of a successful transaction
volume-C	The number of items of product C deposited on the site as a result of a successful transaction
volume-D	The number of items of product D deposited on the site as a result of a successful transaction
	Trader-specific variables
product-A	The number of items of product A the trader owns and can trade or store in this turn
product-B	The number of items of product B the trader owns and can trade or store in this turn
product-C	The number of items of product C the trader owns and can trade or store in this turn
product-D	The number of items of product D the trader owns and can trade or store in this turn
stock-A	The number of items of product A the trader puts in his stock in this turn as a result of an unsuccessful transaction or for redistribution in the next turn
stock-B	The number of items of product B the trader puts in his stock in this turn as a result of an unsuccessful transaction or for redistribution in the next turn
stock-C	The number of items of product C the trader puts in his stock in this turn as a result of an unsuccessful transaction or for redistribution in the next turn
stock-D	The number of items of product D the trader puts in his stock in this turn as a result of an unsuccessful transaction or for redistribution in the next turn
maximum-stock-size	The number of items the trader is willing to obtain through trade this turn in addition to his own demand if the average demand is higher than his demand
price	The price the trader believes an item is worth based on his knowledge of supply and demand on the market
demand	The proportion of the demand at the market the trader is located at that he aims to satisfy by obtaining products through trade. Constant increase of 1 per turn; maximum = *max-demand*

determines the manner in which the remaining traders (or all traders if none were specifically distributed to production sites) are distributed. If this variable is set to 'uniform', the traders are distributed equally among the sites. If this variable is set to 'exponential', the distribution follows an exponential frequency distribution with its mean equal to the amount of undistributed traders (www.comses.net, a). These two sub-models make up the first part of the initialisation.

The second part of the initialisation consists of creating the network of links between traders. The creation of this network is dictated by the *network-structure*, *maximum-degree*, *proportion-inter-site-links*, *proportion-intra-site-links* and *proportion-mutual-neighbors* independent variables. If *network-structure* is set to 'hypothesis', the following steps will be performed. Firstly, a random trader on each site is linked to another random trader on the next site in the circle, so that the whole circle has a minimum level of connectivity. Secondly, inter-site links are created. The amount of trader pairs that will be linked during this step is equal to the total amount of possible trader pairs times the *proportion-inter-site-links* variable. Traders will only be linked if they are not located on the same site, are not already linked and if they have not yet reached the *maximum-degree* of connections. The total amount of possible trader pairs is determined by the following formula, where *n* is the total amount of traders:

$$\frac{1}{2}n(n-1)$$

Note that, in some experiments, *proportion-inter-site-links* is equal to 0 (tab. 1), which means there will be no inter-site links created, other than the ones to connect the circle. Thirdly, an amount of traders equal to the *proportion-intra-site-links* variable times the total number of possible trader pairs are linked if they are located on the same site, are not linked yet and if they have not yet reached the *maximum-degree* of links. Fourthly, traders on the same site with mutual neighbours are linked. A random amount of traders will be selected equal to the number of trader pairs with mutual neighbours times the *proportion-mutual-neighbors* variable. If the selected trader is connected to two or more other traders on the same site, one pair of those will be linked if they are not yet linked to each other and if neither has reached the *maximum-degree* of links. The number of trader pairs with mutual neighbours is calculated with the following formula, where "z_i *is the degree of the* i^{th}*trader*" (Brughmans and Poblome 2016a), in other words z_i is the number of connections a trader has:

$$\frac{1}{2}\sum_{i} z_i(z_i - 1)$$

The last two steps, the creation of intra-site links and the connecting of mutual neighbours on the same site, will be repeated until the average amount of links of all traders, the average degree, reaches the *maximum-degree* minus 10%. According to Brughmans and Poblome (2016a), the repetition of steps three and four will result in a 'small-world' network as presented by Jin *et al*. (2001). In a 'small-world' network most nodes, traders in the case of MERCURY, are not connected directly to each other, but they are indirectly connected by a small number of steps through other nodes. In addition, if a node is connected to two other nodes, those two other nodes have a high chance of also being connected to each other. In other words, a 'small-world' network has a high amount of clusters (Watts and Strogatz 1998, 440). Lastly, if there are multiple clusters, these clusters are connected by creating a link to a trader in another cluster on the same site. If this step is skipped, products cannot be traded across the whole network. If *network-structure* is set to 'random', all previous steps are performed, in order to count the number of links that would have been created, then this network is deleted and a number of new trader pairs are connected equal to the amount of links that were deleted (www.comses.net, a). Figure 2 shows two sample views of the MERCURY world, one created using a very low *proportion-inter-site-links* value and one with an intermediate value.

After the initialisation is completed, the traders begin their trading process, which consists of the following actions. These sets of actions are looped 20 000 times, also called *ticks* in NetLogo and ReLogo jargon. Firstly, each trader's *demand* dependent variable is increased by one if it is less than the *max-demand* independent variable. Each trader's *demand* is zero at the start of each simulation. Secondly, the traders reduce each of its four *stock-X* values by 14% and add the amount removed to the corresponding *volume-X* value of each site. This specific percentage

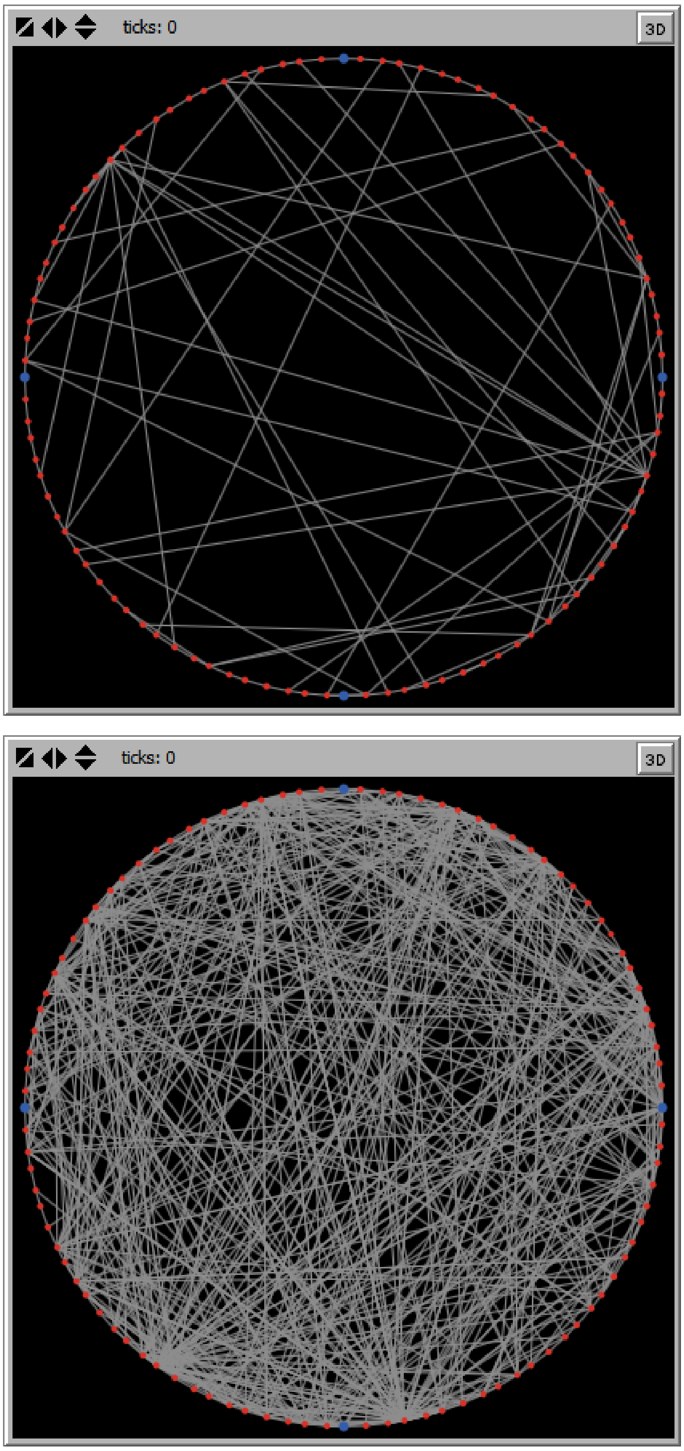

Figure 2: Two example views of the MERCURY world, created using the original NetLogo model (www.comses.net, a). The red dots portray non-production sites and the blue dots portray production sites. The grey lines represent inter-site links. Links between traders on the same site are not visible, since all traders on the same site are occupying the same location. The top image was created using a *proportion-inter-site-links* value of 0,0001 and the bottom one using a value of 0,001.

is based on previous research by Peña (2007, 329). Then the *product-X* values of all traders, which is the part of the trader's product that is tradable, are set to their corresponding *stock-X* values and the *stock-X* values are set to zero. In other words, traders drop 14% of their stock and then their stock becomes tradable during the rest of this loop. The dropping of stock represents the risk of products breaking or becoming out of style when they are not sold to a consumer immediately (Brughmans and Poblome 2016a). Thirdly, all traders located on a production site produce new products by increasing their *product-X* value of the type that is produced at its site by an amount equal to the trader's *demand* minus the sum of all their *product-X* values. Fourthly, traders inform each other on demand and supply. Each trader is assigned a number of randomly chosen informants from the traders they are linked to equal to the amount of traders they are linked to times the *local-knowledge* independent variable. The specific set of informants each trader receives information from changes every *tick*. Then, each trader calculates the average *demand* and average supply of its informants and itself combined. The supply is the sum of all products of every type a trader possesses. This information is used to calculate a price using the following formula:

$$price = \frac{average\ demand}{average\ supply + average\ demand}$$

Fifthly, each trader calculates its *maximum-stock-size*. This value is calculated as the average *demand* of its informants, minus the trader's own *demand*. In other words, only when a trader's average *demand* among its peers and itself is higher than this trader's own *demand*, will it be able to store products that it has obtained from other traders for later loops. Lastly, every product of every trader is traded or considered for trading. This process goes as follows. First, one of the four product types is randomly chosen. Then a random trader who owns any products of the chosen type is selected as the seller. If there are any traders connected to the seller with a *demand* value of higher than zero or a *maximum-stock-size* of higher than zero, these traders are selected as potential buyers. From these potential buyers, the one with the highest *price* estimation is then selected as the buyer. If the buyer's *price* value is equal to or higher than the seller's price, reduce the seller's *product-X* of the type that is being traded by one. In other words, when the buyer believes the product to be more, or at least as, valuable as the other trader, the seller sells the product. If the buyer's demand is higher than zero, decrease its *demand* by one and increase the *volume-X* value of the site that the buyer is located on by one, i.e. the product is consumed and deposited on the site. If the buyer's demand is zero, increase the *stock-X* value of the type of the sold product by one and decrease its *maximum-stock-size* by one; the product is stored to be traded later. If there are no potential buyers, or if the selected buyer's *price* value is less than the seller's *price*, the seller stores all products of that type for later by setting its *stock-X* value of the relevant type equal to the corresponding *product-X* value and by setting its *product-X* value to zero afterwards and decreasing its *maximum-stock-size* by the amount that was added to its stock. This process is repeated until every product of every trader has been considered for trading (www.comses.net, a).

After 20 000 loops have been performed, the results are exported. The exported data consists of all the independent variables, the amount of links that have been created during each of the five steps during the initialisation, the average degree that was reached during the repeating of step three and four of the trader network creation, the clustering coefficient of the trader network, the average shortest path of the trader network and the amount of sites each product type is on, sorted from highest to lowest, not by its original type (Brughmans and Poblome 2016a, supplement 1). The absolute amount of products on each site is not taken into account, just the spread of each product across the network.

2.3 Brughmans and Poblome's results and conclusions

Brughmans and Poblome (2016b, 400-401) use the previously described independent variable *proportion-inter-site-links* to make MERCURY simulations represent Bang and Temin's conceptual models. The *proportion-inter-site-links* variable determines the amount of traders that will be linked between sites, as a proportion of the total amount of possible trader pairs. Brughmans and Poblome (2016a) state that the availability of information was low in both Temin's and Bang's mod-

Table 3: A list of all experiments with their independent variables. This table excludes independent variables that are equal across all experiments (after Brughmans and Poblome 2016a, supplement 1).

Exp.	equal-traders-production-site	traders-distribution	network-structure	local-knowledge	proportion-inter-site-links	traders-production-site	max-demand	proportion-intra-site-links
1	TRUE	exponential	hypothesis	0.1	0	10	10	0.0005
2	TRUE	exponential	hypothesis	1	0	10	10	0.0005
3	TRUE	exponential	hypothesis	0.1	0.0001	10	10	0.0005
4	TRUE	exponential	hypothesis	1	0.0001	10	10	0.0005
5	TRUE	exponential	hypothesis	0.1	0.0006	10	10	0.0005
6	TRUE	exponential	hypothesis	1	0.0006	10	10	0.0005
7	TRUE	exponential	hypothesis	0.1	0.001	10	10	0.0005
8	TRUE	exponential	hypothesis	1	0.001	10	10	0.0005
9	TRUE	exponential	hypothesis	0.1	0.002	10	10	0.0005
10	TRUE	exponential	hypothesis	1	0.002	10	10	0.0005
11	TRUE	exponential	hypothesis	0.1	0.003	10	10	0.0005
12	TRUE	exponential	hypothesis	1	0.003	10	10	0.0005
13	TRUE	exponential	hypothesis	0.5	0.001	1	1	0.0005
14	TRUE	exponential	hypothesis	0.5	0.001	1	10	0.0005
15	TRUE	exponential	hypothesis	0.5	0.001	20	10	0.0005
16	TRUE	exponential	hypothesis	0.5	0.001	30	10	0.0005
17	TRUE	exponential	hypothesis	0.5	0.001	10	1	0.0005
18	TRUE	exponential	hypothesis	0.5	0.001	10	20	0.0005
19	TRUE	exponential	hypothesis	0.5	0.001	10	30	0.0005
20	TRUE	exponential	hypothesis	0.5	0.001	30	30	0.0005
21	FALSE	exponential	hypothesis	0.5	0.0001	na	10	0.0005
22	TRUE	exponential	hypothesis	0.5	0.001	10	10	0.0005
23	FALSE	uniform	hypothesis	0.5	0.001	na	10	0.0005
24	FALSE	exponential	hypothesis	0.5	0.001	na	10	0.0005
25	FALSE	exponential	hypothesis	0.5	0.001	na	30	0.0005
26	FALSE	exponential	hypothesis	0.5	0.002	na	10	0.0005
27	FALSE	exponential	hypothesis	0.5	0.002	na	30	0.0005
28	FALSE	exponential	hypothesis	0.5	0.003	na	10	0.0005
29	FALSE	exponential	random	0.5	0.001	na	10	0.0005
31	FALSE	exponential	hypothesis	1	0.001	na	10	0.0005
32	FALSE	exponential	random	0.5	0.001	na	30	0.0005
33	FALSE	exponential	hypothesis	0.5	0.001	(30,1,1,1)	10	0.0005
34	FALSE	exponential	random	0.5	0.001	(30,1,1,1)	10	0.0005
35	TRUE	exponential	random	0.5	0.001	10	10	0.001

els, which manifests itself in a low *local-knowledge* value, while *proportion-inter-site-links* should be set to high values to reflect the heavily integrated markets of Temin's model and low to reflect weak market integration in Bang's model. The experiments are not limited to variations in *proportion-inter-site-links*, but include a wide range of variations in independent variable settings. Table 3 shows all experiments, excluding experiment 30, and the independent variables that are unique to them. Experiment 30 was excluded because it used a test variable, *transport-cost*, which is not discussed in either of the two articles or the ODD. This table does not include the independent variables that are equal across all experiments. A complete version of this table, including summary statistics of the output data, can be found in Brughmans and Poblome (2016a, supplement 1).

Firstly, Brughmans and Poblome (2016a) used the results of experiments 1, 3, 5, 7, 9, 11 and 35 to study the effect *proportion-inter-site-links* has on the network itself, not on the spread of tableware. In these experiments, *proportion-inter-site-*

links is varied between 0, 0,0001, 0,0006, 0,001, 0,002 and 0,003 for hypothesised networks, while other independent variables are kept constant. Additionally, these hypothetical networks were compared to a random network structure, experiment 35. Two network measures, clustering coefficient and average shortest path length were used to compare the results of these experiments (Brughmans and Poblome 2016a). Watts and Strogatz's (1998, 441) concept of local clustering coefficient is used, which is defined as the proportion of links between the neighbours of a node, a trader in the case of MERCURY, divided by the maximum amount of possible links between these neighbours. The mean of the local clustering coefficient among all traders is used as the clustering coefficient in Brughmans and Poblome (2016a). Thus, a network with a low integration between markets on different sites will have a high clustering coefficient and a network wherein markets are highly integrated will have a low clustering coefficient. Average shortest path length is simply defined as "*the average number of steps along the shortest paths for all possible pairs of network nodes*" (Mao and Zhang 2017, 243). It was found that low *proportion-inter-site-links* values resulted in higher clustering coefficients and lower average shortest path lengths and high *proportion-inter-site-links* values resulted in lower clustering coefficients and higher average shortest path lengths. These outcomes correspond to Bang and Temin's models, respectively. In the case of a randomly created network, average shortest path length was low and clustering coefficient was extremely low (Brughmans and Poblome 2016a). These results show us that *proportion-inter-site-links* can indeed be used to represent the differences between Bang and Temin's conceptual models.

Secondly, Brughmans and Poblome (2016a) used experiments 1 to 12 to study the influence of the independent variables *proportion-inter-site-links* and *local-knowledge* on the product distribution. The same values of *proportion-inter-site-links* as above were used. In addition *local-knowledge* was varied between 0,1 and 1. Every combination between these values of the two variables was used, while the other independent variables were kept constant. These experiments showed that when traders have imperfect information within their network, when *local-knowledge* is set to 0,1 instead of 1, all products will spread wider on average. This difference is consistent but slight. Increasing *proportion-inter-site-links*, on the other hand,

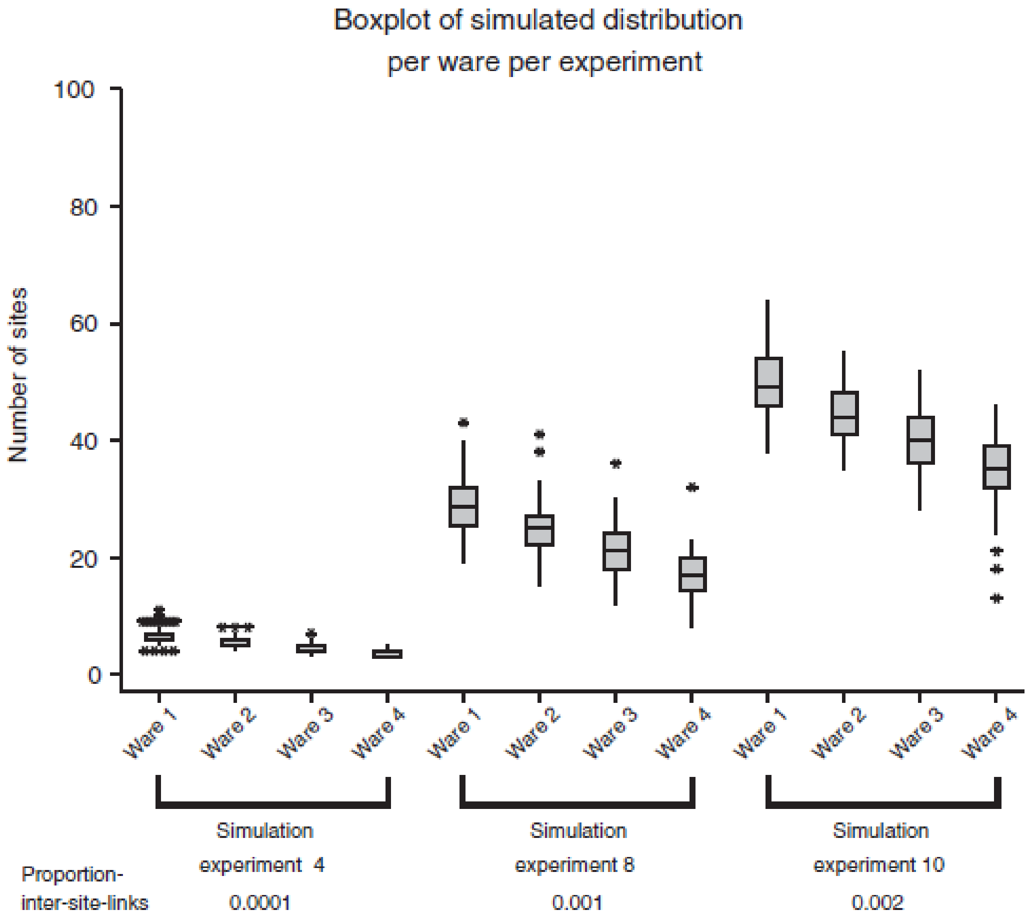

Figure 3: Boxplots of the width of distribution, ranked from most to least widely distributed, from three experiments with different *proportion-inter-site-links* values. All other independent variables are the same. This graph was created from data of 100 iterations per experiment (after Brughmans and Poblome 2016b, 403).

has significant influence on the wideness of ware distributions, the amount of sites each product is found on (fig. 3). However, the difference between the ware with the highest width of distribution and the one with the lowest, called the range of distribution by Brughmans and Poblome (2016a), is low. Therefore, these parameters alone cannot explain the archaeological observations made from the ICRATES data (Brughmans and Poblome 2016b, 401). Thirdly, Brughmans and Poblome (2016a) used experiments 13 to 20 to study the influence of *traders-production-site* and *max-demand* on the product distribution. Combinations of the valuers 1, 10, 20 and 30 for both *traders-production-site* and *max-demand* were used, although not all combinations of these values between the two variables were tried. For these experiments, *proportion-inter-site-links* and *local-knowledge* were set to the moderate values of 0,001 and 0,5, respectively. These experiments showed a similar result as experiments 1 to 12: increasing *traders-production-site* and *max-demand* increases the width of distribution, but does not meaningfully affect the range of distribution.

Fourthly, experiments 21, 23 to 28, 31 and 33 were used to test the influence of setting *equal-traders-distribution-site* to 'false', i.e. distribution to production sites following the same rules as distribution to non-production sites. In addition a uniform *traders-distribution* and an unequal traders distribution to production sites by setting *traders-production-site* to '30,1,1,1', as explained in the previous section, was tested. Other independent variables were not uniform throughout these experiments, for example different values of *proportion-inter-site-links* were tried. Experiments 21 and 24 to 28 showed that increasing *proportion-inter-site-links* increases distribution width, as previously shown in experiments 1 to 12. However, when combining higher values of *proportion-inter-site-links* with *equal-traders-distribution-site* to 'false', a much higher range of distribution is achieved. Setting *traders-production-site* to '30,1,1,1' results in one product, the one whose production site has the highest amount of traders, being spread much wider than the others, i.e. the desired archaeologically observed pattern. Experiment 23, where the distribution of traders among sites was uniform, showed a high width of distribution for all wares, and, consequently, a low range of distribution (Brughmans and Poblome 2016a).

Lastly, Brughmans and Poblome (2016a) used experiments 22, 24, 25, 29, 32, 33, 34 and 35 to compare randomly created networks to hypothetical networks that follow the small-world model. Randomly created networks were compared to several hypothetical networks with varied values for *traders-production-site, maximum-demand* and *equal-traders-production-site* values. This results from these experiments show that all products in randomly created networks spread much more widely than in their hypothetical-network counterparts. A fairly obvious result, since in randomly created networks there are much more trader pairs who are located on different sites from each other, as seen before. But randomly created networks did not have a higher range of distribution. When setting *traders-distribution-site* to '30,1,1,1', the randomly created network shows a higher width of distribution for all

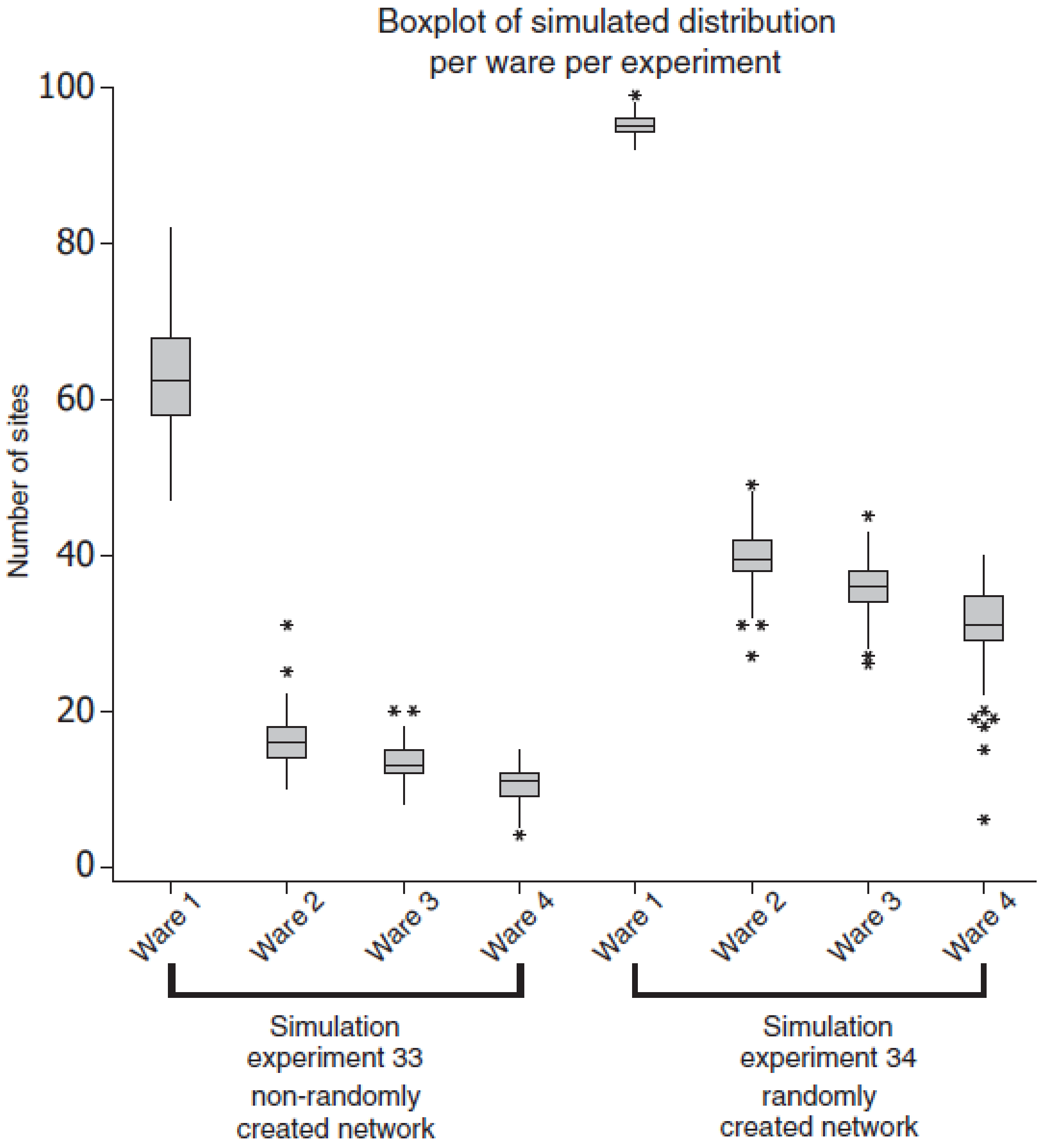

Figure 4: Boxplots of the width of distribution, ranked from most to least widely distributed, from experiments 33 and 34. Both experiments include a disproportional distribution of traders among produciton-sites, i.e. *traders-distribution-site* = ’30,1,1,1’, but the first experiment was created using the hypothesised network and the second using a random network. All other independent variables are the same. This graph was created from data of 100 iterations per experiment (after Brughmans and Poblome 2016b, 405).

products, as well as a higher range of distribution (fig. 4). Brughmans and Poblome (2016a) state that the hypothesised network structure is not as important for explaining the archaeologically perceived ware distribution as placing an unequal amount of traders on production sites.

Brughmans and Poblome (2016a) conclude that increasing the variables *proportion-inter-site-links*, *traders-production-site*, *max-demand* and the creation of randomly created networks, as opposed to hypothetical small-world networks, correspond to increased distribution width, but that they do not give rise to higher ranges of distribution. The only scenario that conform to the distribution patterns perceived in the ICRATES data, is when one production site receives a much higher amount of traders than the other three, since such a site has the ability to export more wares. In addition they claim that: "*The results lead us to conclude that the limited integration of markets proposed by Bang's model is highly unlikely under the conditions imposed in this study. The simulation confirmed the importance of market integration, as suggested by Temin's model, but it also highlighted the strong impact of other factors: differences in the potential production output of tableware production centres, and the demand of their local markets* (Brughmans and Poblome 2016b, 404-405)." Brughmans and Poblome (2016a) also make important remarks concerning the importance of agent-based modelling research and the actions other researchers could take to facilitate further simulation studies.

3 The replication process and its results

This chapter concerns the main body of this thesis: an in-depth explanation of the replication process of MERCURY and the result of that replication. Prior to starting this research, my experience with ABM and coding in general was very limited, and in many ways it still is. The only practice I had with coding was a seven week university course on ABM and an even shorter free online course on programming using Python 2. When choosing which ABM platform to use, a beginner-friendly nature was important to me. The software I decided to use was Repast Simphony. The main reason for this decision was that it employs *ReLogo*, a domain-specific language for ABM with primitives similar to NetLogo (Ozik *et al*. 2013), with which I already had experience. A primitive is a basic element of a programming language that can be used to write code. Repast Simphony also uses *Groovy*, an 'agile' form of Java, partly inspired by Python, meaning it it is generally less verbose and easier to read than Java (König *et al*. 2015, 3-53), which appealed to me. I used several guides from the official GitHub documentation page for practice (repast.github.io, b). The bulk of the programming work was done in 15 days of lab work, during which I also had to get acquainted with Repast Simphony. The ODD, written by Brughmans and Poblome and found at their CoMSES Net / OpenABM page (www.comses.net, a), was used as the main source regarding the specifics of the MERCURY model. Later in the process, I found that one of their articles contains information about the model that the ODD lacks (Brughmans and Poblome 2016a). The original source code was only used if the ODD was lacking, and for comparison after the initial version of the replication was complete. After the first version of the replication was completed, the same experiments as presented in Brughmans and Poblome (2016a) were repeated and compared to the original. If noticeable differences were present, alterations to the code were made and the same process was repeated again. This resulted in eight versions of the replicated model over the course of many weeks. These versions, their results and subsequent alterations will be discussed below. Note that there are minor changes that have no meaningful influence on the output of the model, such as annotation changes and the standardisation of variable names, that will not be discussed here. Even though the ODD was followed literally, there existed many differences between my replication and the original model, some of which influenced the output, as will be discussed later. Some errors in the code were my fault alone and cannot be attributed to the ODD or my interpretation of it.

Before going into the different versions of the replication, I want to make some brief comments about the graphs presented in this chapter. Due to the large amount of data points collected in this study, there are many possibilities for creating graphs. Since the amount of data is so large, presenting it succinctly in a limited amount of graphs is a challenge. Therefore, I urge the reader to consult the appendix if they feel like they are missing crucial information. For the purpose of comparing the network of traders between the replication and the original, the clustering coefficient was used. Unlike most other measures, the clustering coefficient says something about the network as a whole. The average shortest path distance was also an option, but this measure has very large differences between the experiments, even more so than the clustering coefficient does, which made the graphs difficult to read. The four experiments with random networks always have very low clustering coefficients, which makes them difficult to view in all the graphs. Changing the Y-axis to a logarithmic scale, so that a wider range of values can be properly displayed, was considered. However, this option was rejected because it made the differences between the replication and the original less visible, which counteracts the main point of the graphs. Note that in the original study, the network measures, including the clustering coefficient, were only measured for one experiment, the one with a ’random seed’ of 10. A random seed, also simply called a ’seed’, is an input number that determines the output of a pseudo-random number generator (Shamir 1981). In other words, it is a number that determines the ’random’ events of the MERCURY model, so that if the same random seed is used, the ’random’ events will have the same results. For the graphs of the ware distributions, both mean distribution width and range was used, as they are both summary statistics of the ware distribution and say more than other statistics such as the minima, maxima or mode. All original graphs in this thesis were created using LibreOffice Calc (libreoffice.org).

3.1 Version 1

Version 1 of the replication is defined as the first version that included all features described in the ODD, could export all the data as in supplement 1 of Brughmans and Poblome's (2016a) paper in the *Journal of Artificial Societies and Social Simulation*, or JASSS, and could run successfully, without fatal errors. The source code is added as an appendix (appendix 2). My goal was to replicate the MERCURY model using only the ODD as a source. There were, however, three times where the ODD, or my understanding of it, did not suffice, and the source code had to be consulted. Firstly, the term "exponential frequency distribution" in the following passage was unclear to me: "*When equal-traders-production-site is set to "false", all traders are distributed among all sites following a uniform or exponential frequency distribution, depending on the setting of the variable traders-distribution. The mean of the exponential frequency distribution is the number of traders that have not yet been moved to a site divided by the number of sites* (www.comses.net, a)." After looking through the original code, I found out that what was meant by this is that each site has a target distribution, an amount of traders on the site that should be met, that is equal to a random number from an exponential distribution with a mean equal to the amount of traders that are not moved yet, rounded up. Perhaps my knowledge of mathematics was simply not sufficient to understand this usage of the term, so I will leave it up to the reader to judge the clarity of the ODD in this case. The second and third times the source code had to be consulted concerned the reporters that were used to calculate the clustering coefficient and the average shortest path distance. The specific way in which average shortest path distance was calculated was not mentioned in the ODD, nor in the articles (Brughmans and Poblome 2016a; Brughmans and Poblome 2016b). There exist multiple ways to determine the shortest path, such as Dijkstra's algorithm and its variants, or the

Bellman-Ford algorithm, each with different uses (Festa 2006). In the source code, a primitive, *mean-link-path-length*, from a network extension for NetLogo is used (www.comses.net, a). Since the creator of this network extension also does not mention which algorithm is used (github.com), I opted to use the *ShortestPath* Repast package (repast.sourceforge.net, b), which adopts Dijkstra's algorithm. For the calculation of the clustering coefficient, I chose to consult the source code because it was not mentioned in the ODD if the global clustering coefficient was used or the averaged local clustering coefficient. The procedure used to calculate the clustering coefficient was adopted from the original model. Later, I found out that the latter is used in the small-world model by Watts and Strogatz (1998, 441) which the authors of MERCURY reference (Brughmans and Poblome 2016a), so in this case it might not have been entirely necessary to turn to the source code, but it would have been better if it was explained more clearly in the ODD.

The data from the 34 experiments of version 1 of the replication can be found in appendix 3. Originally, Brughmans and Poblome (2016a) described 35 experiments in their supplement table. One of the experiments, number 30, involves the use of a variable named *transport-cost*. A similar variable, *transport-fee* is mentioned in supplement 2, which lists all the variables of MERCURY. However, experiment 30 is not discussed in the articles and neither variable occurs in the code. In one article, the possibility of incorporating transport costs in a future version of MERCURY was mentioned (Brughmans and Poblome 2016a). In an email correspondence with Tom Brughmans, he told me that the variable *transport-cost*, which was used in experiment 30, was a leftover of a testing phase and did not make it in the published version (appendix 19). During the course of writing this thesis, after my email correspondence with Tom Brughmans, an extension of MERCURY which incorporates *transport-cost* into the model. However, the corresponding paper has not been released yet. In all appendix files containing the summarised data from the replication, the original numbering is used, and experiment 30 is skipped, but in the raw data tab, number 30 is not skipped, so experiment 30 is equal to 31 in the summarised table, 31 is equal to 32, etc.

As can be seen in appendix 3 there are major discrepancies between the out-

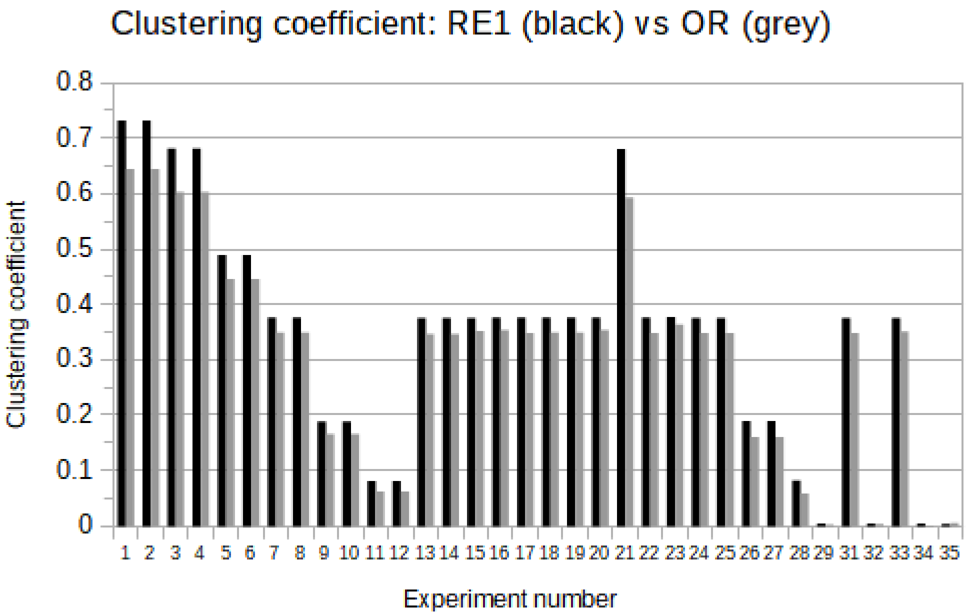

Figure 5: A bar graph showing the mean clustering coefficient of all 100 iterations per experiment of version one of the replication next to the clustering coefficient of seed 10 of the original model, sorted by experiment number.

put data from version 1 of the replicated model and the output data from the original MERCURY model. Appendix 3 contains a table, MERCURY_pct_change, that shows the percentage change from the data of the original model to the replicated model. It should be noted that the network measures of the original model were only recorded on one run per 100 runs for each experiment. Therefore, the table shows a comparison between the averaged network measures from the replicated model and the network measures from one run of the original. This only applies to the network measures; the ware distribution data was averaged for every run in the experiment in the original study. The percentage change for the links created in step one and two, the linking of one trader per site on the circle and the creation of inter-site links, is 0,00%. This means that the amount of links created in these steps is identical for each run per experiment and that this number is equal to the one from the original, as should be the case. The amount of randomly created intra-site links, on the other hand, is consistently higher in the original: ranging from a change of -11,06% to -4,66% from the original to the replication. On the contrary, the number of intra-site links created through mutual neighbours is consistently lower in the original, ranging from change of 4,59% to 33,09%. These two steps are performed in a loop until an *average-degree* of 4,5 is met. It would seem that during this loop, one of these processes creates either less or more links than intended, which results in a skew in the other process as well. The amount of links created to connect components varies wildly, from percentage change of -45,21% to 532,00%, although in the majority of cases, 22 out of 34, it is lower in the original. These discrepancies seem to cancel each other out, as the total number of links is very similar between the two datasets, ranging from -1,03% to 0,11%. This also holds true for the average degree, which ranges from -0,55% to 0,09%. The clustering coefficient is consistently higher in the replication, ranging from a change of 3,69% to 52,83%, compared to the original, as can be seen in figure 5. The last two experiments, 34 and 35, are major exceptions with values of 279,30% and, the only experiment with a higher clustering coefficient in the original version, -29,41% change. Average shortest path distance ranges from -1,01% to 4,08% change, except for the first two experiments with extreme values of -49,20% and experiment 20 with 17,62%. These percentage changes between the original and the replication clearly show a difference in network creation. In future versions, if the data is more similar, I will use different ways of comparing this data, but looking at the percentage change suffices for the time being.

In terms of ware distribution measures, there are major discrepancies as well. In general, all ware distribution measures are much higher in the original. I will not go into detail about the ware distribution data here, because I decided to try and correct the network generation process first, as this influences the ware distribution, while the opposite is not the case.

3.2 Version 2

Version 2 of the replication includes a minor change to the loop that repeats the creation of random intra-site links and links between mutual neighbours and a major change in the code that dictates the creation of random intra-site links. The source code of this version of the replication can be found in appendix 4.

Firstly, the while loop that dictates the creation of random intra-site links and mutual neighbour intra-site links was changed so that the average degree has to be less than or equal to the *maximum-degree* minus 10%, instead of just less than it. In the ODD, this process is described as follows: "*Steps three and four of this network creation procedure are repeated while the average degree of the network is lower than maximum-degree minus 10%* (www.comses.net, a)." Because of the words "lower than", I chose to use a less-than operator instead of the less-than-or-equals-to operator which was used in the original code. Nevertheless, this is a minor change that results in a very small difference in the number of links that are created.

Secondly, the way in which conditions are checked in the creation of random intra-site links was changed. In the ODD, the pairing of randomly selected traders on the same site is described as follows: "*Thirdly, randomly selected pairs of traders on the same site are connected. More formally, a proportion of all trader pairs determined by the variable proportion-intra-site-links are connected if they meet the following requirements: both are located at the same site, the pair is not connected yet, and neither of the traders has the maximum-degree* (www.comses.net, a)." My interpretation of this was that all conditions have to be checked after a potential pair of traders is selected. However, this is not consistent with the source code, where the condition of the second trader in the pair having to be located on the same site as the other one, is a requirement for this trader to be selected in the first place. Additionally, the other requirements are checked throughout the process, i.e. whether the first trader has already reached the maximum degree of links is checked immediately after it is selected, instead of after the second trader is selected. In my view, the description in the ODD is ambiguous, as it does not differentiate between conditions that have to be tested just before a link is created and requirements for the selection of trader pairs. The explanation is phrased in a way that suggests all conditions have to be checked simultaneously after a pair is selected. The timing of checking conditions should not matter in this case in terms of results, it might decrease processing time though. However, a condition being a requirement for selection instead of a condition that is checked at the end, can make a difference, as only a limited amount of pairs are selected every time this piece of code is looped, so it will result in less links being created.

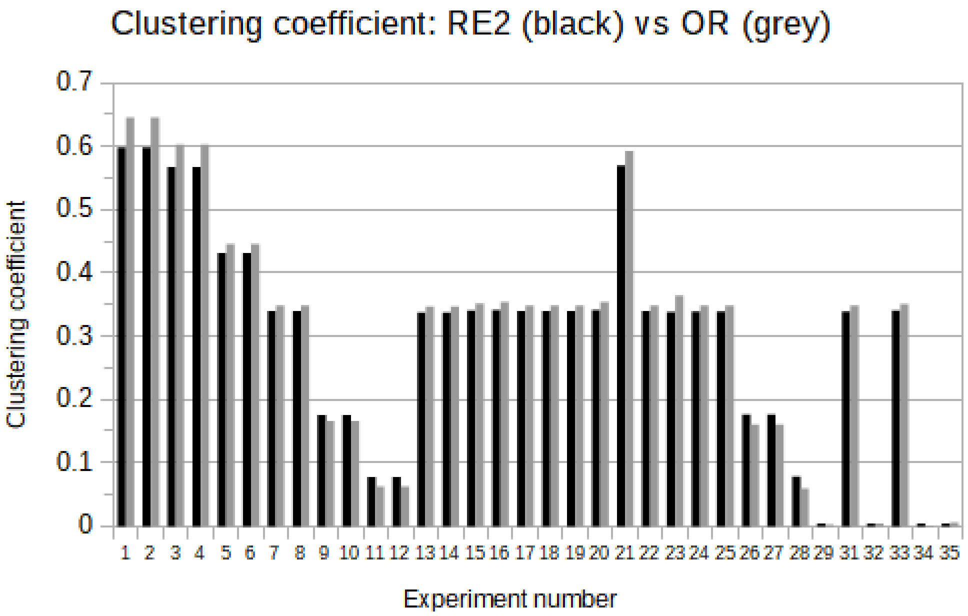

Figure 6: A bar graph showing the mean clustering coefficient of all 100 iterations per experiment of version two of the replication next to the clustering coefficient of seed 10 of the original model, sorted by experiment number.

The data from the 34 experiments of version 2 of the replication can be found in appendix 5. The ratio of randomly created intra-site links and links created between mutual neighbours is skewed towards the former, as opposed to a skew towards the latter in version 1 of the replication. The amount of randomly created intra-site links across all experiments has increased by 10,96% and the amount of links created between mutual neighbours has decreased by 10,23%, compared to version 1. I believe this difference is caused by the second change made in this version. Because there are only a certain amount of pairs that are selected each time the procedure of creating random intra-site links is run, the amount of pairs that will be created each time will increase if one of the conditions for creation is a requirement for the selection of trader pairs. The steps of random intra-site link creation and the linking of mutual neighbours are looped together until the required average degree is reached, so if the amount of links created by one of these procedures increases, the other decreases. The same general patterns of irregularities, as compared to the original, exist in the other network measures of version 2 as they do in version

1. The clustering coefficient is now generally lower in the replication than in the original, instead of smaller. This can be clearly seen by comparing figure 6 to figure 5. Compared to version 1, the clustering coefficient has decreased by 11,92%, but the same outliers in clustering coefficients are still present. The extreme differences in ware distribution data also still exist in the second version. For now, I will continue to make changes to the network creation to make it match with the original. In particular, the creation links between mutual neighbours will be looked into, because it's counterpart, the pairing of randomly selected intra-site neighbours, has already been checked and the ratio between the two is not yet in agreement with the original model's data.

3.3 Version 3

The third version of the replication contains a major overhaul of the pairing of mutual neighbours. The source code of this version of the replication can be found in appendix 6.

There were two ways in which the code of the previous version of the replicated model differed from the original, both having to do with the selection of traders that will be connected. The linking of mutual neighbours is described in the ODD as follows: "*[A] number of traders are selected uniformly at random; the number of selected traders is a proportion of all trader pairs with a mutual neighbour (the proportion is determined by the variable proportion-mutual-neighbors, and the number of trader pairs with a mutual neighbour is calculated as the equation below); if these randomly selected traders are connected to a pair of traders on the same site that are not connected yet and do not have the maximum-degree, then such a pair of traders of whom the randomly selected trader is a mutual neighbour will be connected* (www.comses.net, a)." For the formula in question, see page 23. I interpreted this section in the following manner. Firstly, select a number of traders equal to the total amount of traders with mutual pairs times the *proportion-mutual-neighbors* variable. Then, every one of these traders' neighbours randomly selects another one of the initially selected trader's neighbours, until a pair is selected that is not linked yet, are on the same site and have both not yet reached the maximum degree of links, in which case a link is created between this pair. A variable was included that is used to track if a link has been created yet, so that no more than one pair is connected per initially selected trader. This interpretation differs from the source code of the original model in two ways. Firstly, and most importantly, in the original code, the selection of traders is not uniform, contrary to the explanation in the ODD. Instead, a list is created wherein each trader is added a number of times equal to $z_i(z_i - 1)$, from which traders are randomly chosen. In this calculation, z_i is the number of connections the trader has, also called the degree of trader *i*. Therefore, the selection of traders is not uniform, but weighted towards traders with more neighbours. It should be noted that this part of the procedure in the original code does agree with the model by Jin *et al*. (2001, 6) on which it is based. I later found out that this proportional probability is described elsewhere in the ODD, in the 'stochasticity' section. It was not mentioned in the submodels section, where it should be explained in detail (Grimm *et al*. 2006, 119). Secondly, instead of selecting a specific amount of traders, whose neighbours are then asked to create a link between them, the procedure is repeated a number of times equal to this number. Each time it is repeated a random trader from the list is selected. This way, the same trader could be selected twice, which would not be the case if a number of traders would be selected at once.

The data from version 3 of the replication can be found in appendix 7. Compared to version 2, the relevant network measures are closer to the original. The amount of randomly created intra-site links has decreased and the amount of links between mutual neighbours has decreased. Both numbers now match the original more closely. The percentage change of intra-site links and mutual neighbour links from the original NetLogo model to version 3 of the replication range from -3,37% to 1,87% and -1,50% to 14,65%, which is a big outlier. For version 2, these values were -5,32% to 9,38% and -7,30% to 21,99%. However, there still exists a clear difference in the replication compared to the original in terms of network creation, as can be seen by the clustering coefficient, which is now consistently higher again in the replication instead of lower, as it was in version 1 (fig. 7). Since the intra-site links are now fairly equal to the original, this means that another link creation

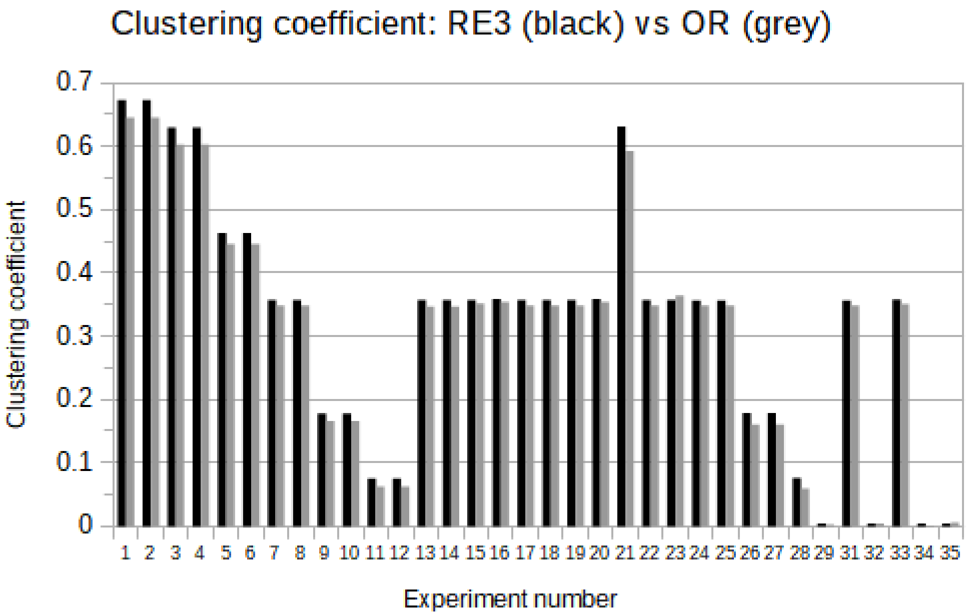

Figure 7: A bar graph showing the mean clustering coefficient of all 100 iterations per experiment of version three of the replication next to the clustering coefficient of seed 10 of the original model, sorted by experiment number.

process, possibly the linking of components, distorts the clustering coefficient. In terms of ware distribution, nothing has changed noticeably; the same massive differences between the original and the replication exist in version 3. For version 4, I will continue to check the network creation for faults and compare it to the original NetLogo code.

3.4 Version 4

Version 4 of the replication includes more error fixes in the network creation. These errors have several causes: my own blunder, an incompleteness in the ODD and an unexpected difference between NetLogo and ReLogo primitives. In addition, an error in the dropping of tableware was fixed. While comparing the code, an error in the original NetLogo code was found. The source code of this version of the replication can be found in appendix 8.

As mentioned in the last section, at this point in the replication process I was still focused on fixing the network creation. However, while going over this code I coincidentally also noticed an error in the tableware distribution part of the model that I decided to fix. During every loop, all traders drop 14% of their stock onto the site that they're located, which "*reflects the risks involved in not immediately selling an item on to a consumer but storing it for redistribution and represents broken or unfashionable items*" (Brughmans and Poblome 2016a). Traders store their wares in a *product* and a *stock* variable, see the dependant variables table on page 21. By accident, I programmed it so that traders drop 14% of their *product* instead of their *stock*. At this point in the loop the *product* variable is always 0, as only after dropping a part of their *stock* is the remaining *stock* transformed into the tradable *product* variable.

While comparing the original code to the replication, I noticed the equal spacing between production sites along the circle. This requirement is not mentioned in the ODD, but it is mentioned in one of the papers (Brughmans and Poblome 2016a). The code of the replication was edited to conform with the original NetLogo code.

Running the model with these previous changes to code in place resulted in a fatal error during the setup of the network of a specific run. The alteration to the dropping of tableware only takes effect during the trading loop, after the network has been created, so I figured the latter alteration, the creation of equal spacing between production sites, must have been what caused the error to occur. The error occurred during the procedure that connects one trader per site in the circle and only occurred during experiment 21 with the random seed 11. The error was caused because the *oneOf()* ReLogo primitive, which selects one random item from a list or set of agents, was called on an empty list, the *trader_list*. The *trader_list* is a site-specific variable that contains all traders that are located on the site in question. It is not required by the ODD, but I created it for ease of use. This is possible because the traders do not move to other sites during the trading process. What this means is that there was one site which did not have any traders assigned to it, which should not be the case. The independent variables that dictate network creation for experiment 21 are *network_structure* == 'hypothesis', *equal_traders_production_site* == 'false', *traders_distribution* == 'exponential'. While inspecting the corresponding code, I noticed two faults in my code. Firstly, instead of checking whether all sites have met their target distribution yet, it only checks the *non_producer_list*, i.e. all sites that are not production sites. I suspect this fault was caused by having copied over this code from the procedure that dictates unequal allocation to production sites compared to non-production sites without editing the relevant part. Secondly, it initially distributes the traders only to non-production sites. Only after the target distributions of the sites have been met, are traders distributed to all sites including production sites. Because *equal_traders_production_site* is set to 'false', all traders should be distributed among all sites. Interestingly enough, when comparing the updated replication's code to the original, I noticed that a similar mistake is present in the NetLogo code. In the NetLogo code, after the target distribution is met, the remaining traders are only distributed to non-production sites. The ODD does not mention this, so I assume this is a fault in the NetLogo code. The amount of traders that are distributed this way is small, so this error is not of major significance. The relevant parts of the code were fixed accordingly, and the bug that was present in the NetLogo code was retroactively added to the creation of uniform trader distributions as well as exponential, as the bug in the original NetLogo code was present in both forms of network creation.

While trying to figure out what caused the previously mentioned error, a difference between the *randomExponential()* ReLogo primitive and NetLogo's *random-exponential* primitive was found. In MERCURY, this primitive is used to calculate a target amount of traders a site has to obtain during network generation when an exponentially distributed network is used. In NetLogo's documentation, the primitive is described as follows: "*random-exponential reports an exponentially distributed random floating point number. It is equivalent to (- mean) * ln random-float 1.0.*" (ccl.northwestern.edu, b). In ReLogo's documentation it is described in a comparable way: "*Returns a random floating point number (exponentially distributed).* ***param****: mean a number* ***return****: random floating point number (exponentially distributed with mean mean)*" (repast.sourceforge.net, a). The similar descriptions, combined with the fact that ReLogo is partially based on NetLogo (Ozik *et al*. 2013, 1560), let me to assume that these primitives would function the same way across both platforms. In this revision, I found out that Repast's primitive's output was consistently lower, and always smaller than one. Because the target distribution is rounded up, this resulted in a target distribution of one for all sites. Not being very familiar with exponential distributions, I simply opted to replicate the previously cited formula that NetLogo's random exponential distribution primitive employs. Later, I read up more about exponential distributions and did some more tests to compare the two primitives, which lead me to discover that Repast's primitive uses the inverse of the parameter as the mean of the distribution, while NetLogo's uses the parameter itself as the mean.

As to why the error was brought about by a change in the spacing between production sites, I am not entirely certain, but my suggestion is as follows. The site in question that did not receive any traders was a production site, which means that with one of the previously mentioned bugs still intact, its trader count was not checked to be higher than it's target distribution and traders were not distributed to it to meet its target distribution. Normally, this should cause an error in a much higher percentage of cases, as only a few traders are randomly distributed after target distributions have been met. However, the fault in the exponential distribution of traders caused the target distribution of all sites to be only one, which means that all traders, minus the number of non-production sites, i.e. 904 traders, were randomly distributed. The chance that a production site, which until that point has not received a trader yet, would not receive any of the 904 randomly distributed traders is quite low. This is why the error only occurred in a very specific situation, during a run with a specific random seed of experiment 21. The other runs of experiment 21 were not affected. Therefore, I believe that the error was not directly caused by the change in spacing along the circle, but that this change caused the random number generator to generate different random numbers for procedures that follow the altered production site allocation than in the previous versions. This caused the random distribution of the 904 traders to be different from version 3 of the replication, which in the specific case of experiment 21 and seed 11, caused one production site not to receive any traders.

In addition to these fixes, a local variable *site_list* was created for the part of the procedure that dictates trader distribution when *equal_traders_production_site* is set to 'false'. This variable contains all the sites to which no traders have been moved yet. When *equal_traders_production_site* is set to 'false', traders are distributed to sites in this list, instead of distributing them to all sites directly. For most

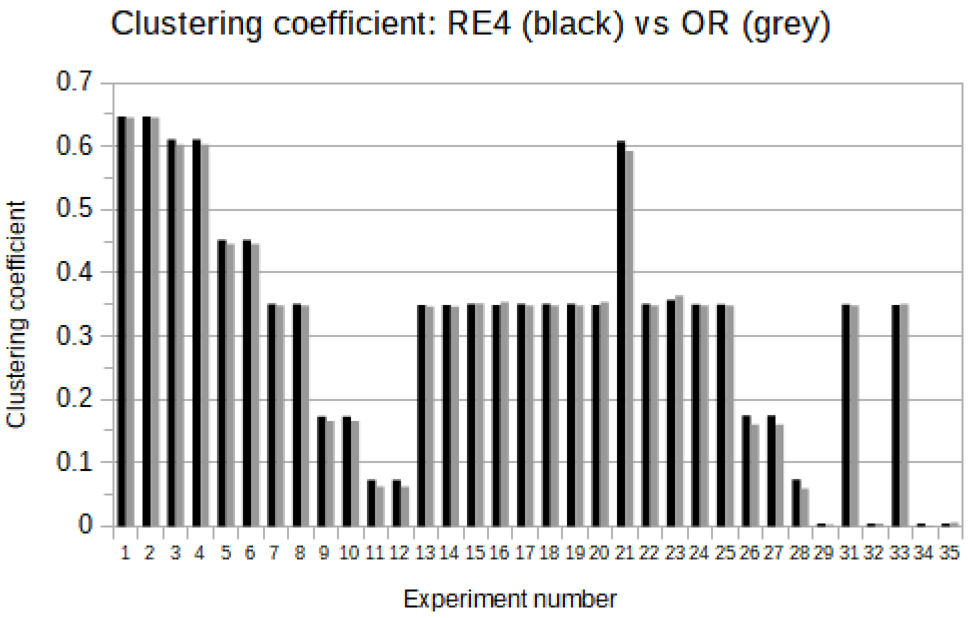

Figure 8: A bar graph showing the mean clustering coefficient of all 100 iterations per experiment of version four of the replication next to the clustering coefficient of seed 10 of the original model, sorted by experiment number.

experiments, this makes no difference, as this list will simply contain all sites. However, this change is needed for experiments 33 and 34, the two experiments where one production site receives 30 traders while the others receive only one. In these experiments *equal_traders_production_site* is set to ’false’, but the production sites should receive a fixed amount of traders, just not an equal number. In previous versions of the replication, the productions sites would receive additional traders if they had not yet reached the target distribution, which would result in a more even distribution between the production site that initially received thirty traders and the other ones. I believe this was not the intention of these two experiments, although the specifics of it are not mentioned in the ODD, the source code or in the articles (Brughmans and Poblome 2016a; Brughmans and Poblome 2016b), so I am unable to confirm this assumption. Excluding the production sites from the *site_list* eliminates this problem.

The data from the 34 experiments of version 4 of the replication can be found in appendix 9. Compared to version 3, the difference in clustering coefficient between the replication and the original model has lessened. The mean percentage change in clustering coefficient from the original to the replication across all experiments was 13,91% in version 3 and 11,86% in version 4. The median has changed from 3,44% in version 3 to 0,72%. This change can be clearly seen in the graphs; the pairs of clustering coefficients in figure 8 are much closer to each other than in previous versions. However, even if the differences are quite small now, the clustering coefficient is still consistently higher in the replication. Only six out of 34 experiments have a lower clustering coefficient in the replication. Therefore, I suspect there is still a difference between the replication and the original and I will continue to review the network creation process for the next version. In terms of ware distribution, there are still massive differences between the original and the replication, however, the amount of experiments with a minimum distribution width of 0 has decreased.

3.5 Version 5

For version 5 of the replication, the setup procedure was checked another time in its totality. This resulted in a change to the creation of inter-site links, a complete overhaul of the *connect_components()* procedure and a change to the creation of random networks. The code for version 5 of the replication can be found in appendix 10.

When creating inter-site links during the second step of the network creation, it was previously not checked whether the traders to be connected had reached their maximum degree of links. This procedure is described in the ODD as follows: "*A proportion (determined by the variable proportion-inter-site-links) of all trader pairs are connected if a pair is not located on the same site and is not connected yet* (www.comses.net, a)." The requirement to check for the maximum degree of links is not mentioned in the ODD as it is in other cases, however, it could have been assumed as the maximum degree is a universal limit. In addition to this change, the other requirements were changed so that the pair of traders are selected for them, instead of the requirements being checked after a pair has been selected, but before a link is created. This change is similar to a change to the creation of intra-site links in version 2, with the same ambiguity in the ODD. In this case, the timing of checking the requirements does not matter in terms of the amount of links that are created, because the procedure is repeated until a predetermined amount of links are created. However, the change was still made because it might decrease run time, as this will reduce the amount of loops that have to be performed.

The *connect_components()* procedure of the replication was found to be entirely incorrect. I misinterpreted the ODD and thought that the purpose of this procedure was to link all traders that were not yet linked to other traders in order to incorporate them into the network. Because I programmed this function incorrectly, there existed multiple clusters of traders which were not linked together. The actual purpose of this procedure is to identify these separate clusters within the network, the components, and connect them to each other so that the network consists of one large component. A procedure was added that identifies all clusters of traders in the network and the already existing *connect_component()* procedure was completely altered so that it connects these clusters to each other instead of just the individual traders without links. The *WeakComponentClusterer* algorithm from JUNG, Java Universal Network/Graph Framework, was used to identify clusters in the network (jung.sourceforge.net). The *connect_component()* procedure differs slightly from its original NetLogo counterpart; instead of randomly selecting a site and connecting two traders on it that belong to different components, a site is selected from a list of sites that have traders that belong to different components on them. This greatly reduces the amount of times the while loop has to be run, as there would be a big chance that a randomly selected site does not have traders from different clusters on it, in which case the while loop repeats itself.

The creation of random networks was changed to include the connecting of components after random links are created. The ODD does not mention that the *connect_components()* procedure has to run after a random network is created, however, it is included in the original code.

The data from the 34 experiments of version 5 of the replication can be found in appendix 11. The average network measures of the replication are very similar to the original's in this version, except for the links created to connect components

Table 4: This table shows whether the network measures of seed 10 from the original study (Brughmans and Poblome 2016a) are greater than or equal to the minimum and smaller than or equal to the maximum of the same network measures of version 5 of the replication. Note that this does not include the first two steps of network creation, as they always create the same amount of links in both the replication and the original. Link amounts for step three and four were not counted in experiments with randomly created networks in the original, which results in missing data: ’na’.

Exp.	Step 3 links	Step 4 links	Step 5 links	Total links	Avg. degree	Clustering coefficient	Avg. shortest path length
1	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
2	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
5	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
6	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
7	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
8	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
9	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
10	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
11	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
12	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
13	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
14	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
15	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
16	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
17	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
18	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
19	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
20	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
21	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE
22	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
23	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
24	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
25	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
26	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
27	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
28	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
29	na	na	na	TRUE	TRUE	TRUE	TRUE
31	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
32	na	na	na	TRUE	TRUE	TRUE	TRUE
33	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
34	na	na	na	TRUE	TRUE	FALSE	TRUE
35	na	na	na	TRUE	TRUE	TRUE	TRUE

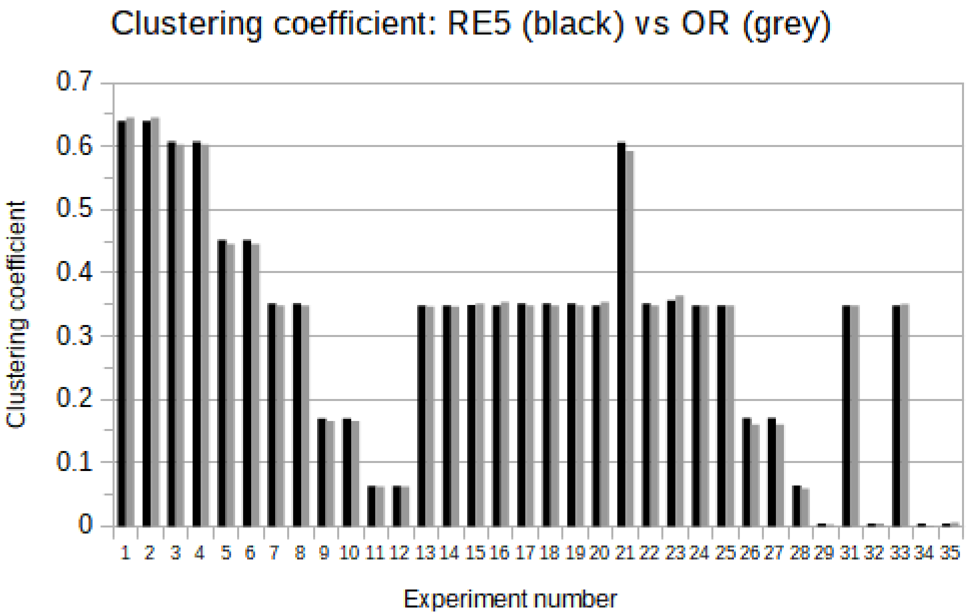

Figure 9: A bar graph showing the mean clustering coefficient of all 100 iterations per experiment of version five of the replication next to the clustering coefficient of seed 10 of the original model, sorted by experiment number.

because they can vary hugely and they're only recorded once for every experiment in the original study. The same outliers still exist, namely the clustering coefficients of experiments 29, 34 and 35 and the average shortest path length of experiment 21, however, these could be the result of the same issue: their being recorded only once in the original. If you compare the clustering coefficients of version five of the replication (fig. 9) to the previous version (fig. 8), there does not seem to be a big difference. I believe this is because the most important change in version five, the overhaul of the *connect_components()* procedure, did not result in a significant change in the amount of links created, since this procedure creates a small amount of links regardless, but rather where in the network they are created. Even though the clustering coefficient is close to the original at this point, it is still generally greater in the replication, however the number of experiments in which it is smaller in the experiment has risen from six to nine. Yet again, it is possible that the fact that this measure is generally greater in the replication is due to the limited data from the original model; the clustering coefficient was only reported for one iteration of each experiment. At this point in the replication process the entirety of network creation has been checked and compared. Therefore, I will move on to a more adequate comparison of the network measure.

There are no statistical methods to compare a distribution to a single measurement. For that reason, I believe that the best way to test whether the network measures of the original match with the replication's is to check whether the original's is higher than or equal to the minimum and lower than or equal to the maximum of each output measurement of the replication. The results of this test can be seen in table 3 on page 50 and the complete table, which includes all the minima and maxima of the network measures of the replication and the original which were used to generate table 3, can be found in appendix 1. There are only two cases in which the original doesn't fall between the replication's minimum and maximum: for the clustering coefficient of experiment 34 and for the average shortest path length of experiment 21. The specific code which dictates network creation for the former experiment cannot be compared to the original, as it uses '30,1,1,1' for *traders-production-site*, which is not included in the version of MERCURY published on CoMSES Net / OpenABM. In the latter case, the average shortest path of the original, 13,2599, is only slightly lower than that of the replication, which ranges from a minimum of 13,3345 to a maximum of 16,3758. It is possible that this is a result of the stochastic nature of the model; with the data available to me, it is not possible to investigate it further.

After the completion of version 5, the network creation part of the code was checked and compared to the original multiple times again in its totality and no errors that could create outcomes in these experiments were found. One very minor difference was found, which I will get to in the next section.

Since the the entirety of the network creation process has been looked into, the ware distribution part of the code will be assessed in the next version. For comparison with later versions of the replication, graphs of the mean distribution width and range of version five are presented here. Both the the mean distribution width (fig. 10) as well as the mean distribution range (fig. 11) is much lower in version five of the replication than in the original model, as they were in previous versions. An

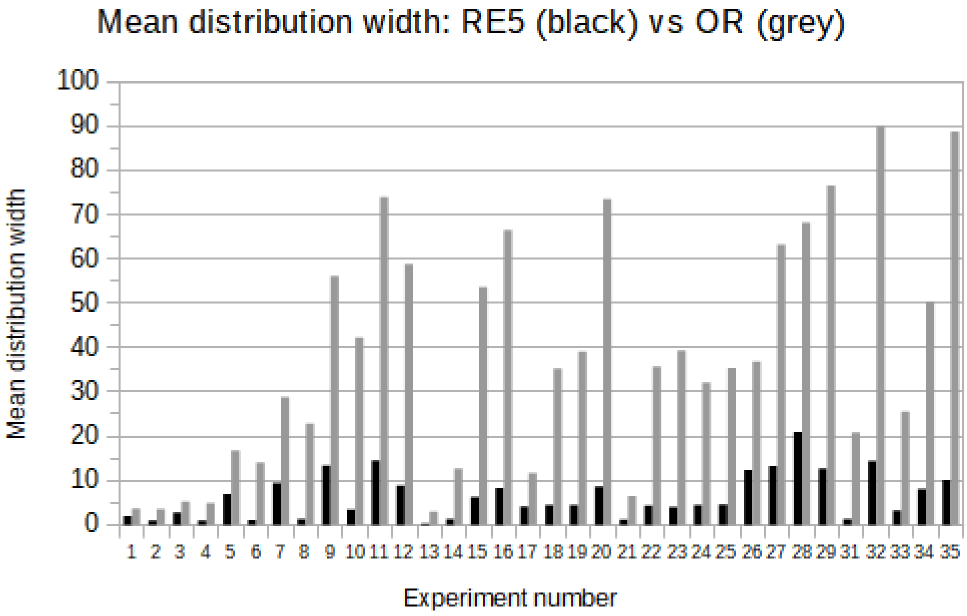

Figure 10: A bar graph showing the mean distribution width of all 100 iterations per experiment from version five of the replication next to the same measure from the original model, sorted by experiment number.

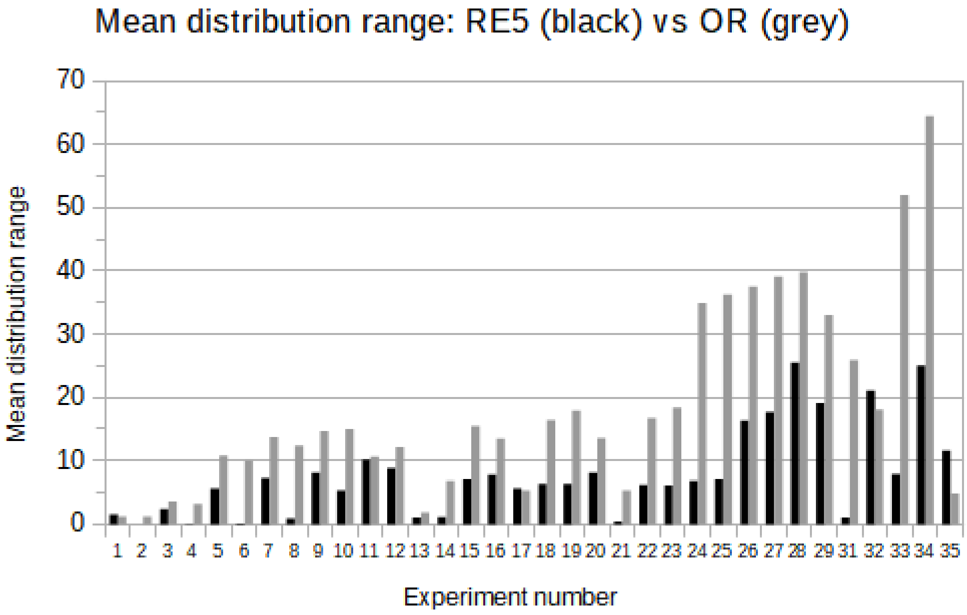

Figure 11: A bar graph showing the mean distribution range of all 100 iterations per experiment from version five of the replication next to the same measure from the original model, sorted by experiment number.

interesting pattern is that for the first twelve sets of experiments, which are pairs of experiments with differing *proportion-inter-site-links* between pairs and a *local-knowledge* of 0,1 or 1 within pairs (tab. 3), the median distribution width is much lower in the experiments with a low *local-knowledge.* For the first eight experiments this is so extreme that their median distribution width is only 1. In the original study, the experiments with a high *local-knowledge* also show a lower distribution width than their counterparts with a low *local-knowledge*, but the difference is not nearly as extreme (Brughmans and Poblome 2016a).

3.6 Version 6

Version 6 of the replication includes changes to the selection of informants, the determination of maximum stock size, as well as an aforementioned inconsequential change to the creation of inter-site links. The code for version 6 of the replication can be found in appendix 12.

Firstly, in the original model, the amount of inter-site links that are created during network creation is a proportion of total pairs, determined by the *proportion-inter-site-links*, rounded *up*, while in the replication, this is rounded to the nearest integer. In practice, this does not affect the simulation at all as in all cases the fraction of the proportion of the total amount of pairs is higher than *x+0,5*, and thus it is always rounded up. Even if it would make a difference, it could only ever result in one less link being created. Nevertheless, the replication was altered. The requirement to round up or not is not mentioned in the ODD.

Secondly, a similar problem was detected and fixed in the code that determines the amount of informants that have to be selected when traders estimate the price of their products. In the original code, the amount of informants to be selected is determined as the number of neighbours the trader in question has times the *local-knowledge* variable, rounded *up*. Again, in the replication, this proportion was rounded to its nearest integer. Rounding this fraction up can result in meaningful differences when *local-knowledge* is set to 0,1; if a trader has fewer than five connections, the product of their number of connections and the *local-knowledge* variable would be rounded down to 0, instead of up to 1. In the experiments with *local-*

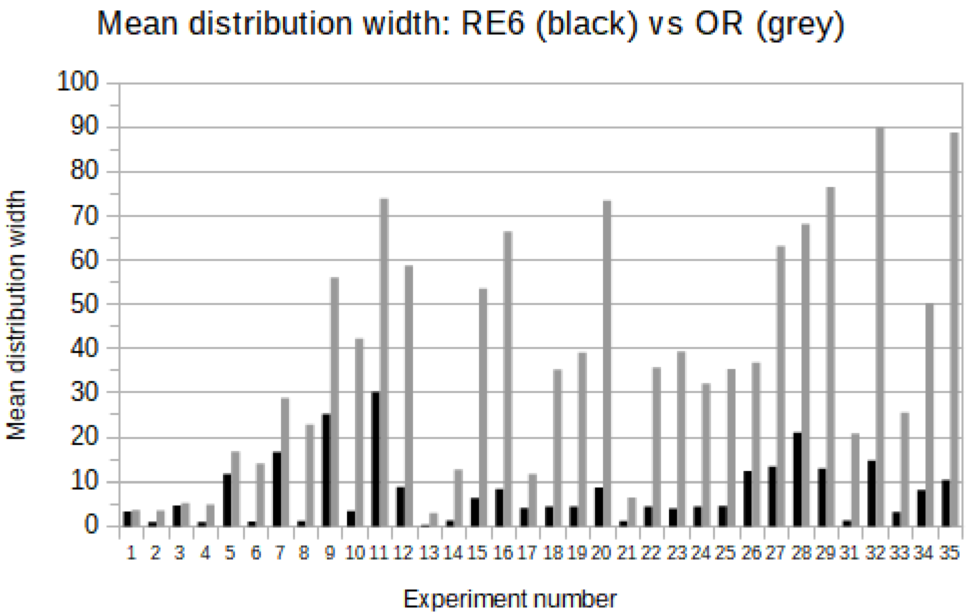

Figure 12: A bar graph showing the mean distribution width of all 100 iterations per experiment from version six of the replication next to the same measure from the original model, sorted by experiment number.

knowledge = 0,5, it would be rounded up regardless, and when *local-knowledge* = 1 there would be no rounding. Whether to round this figure up or down was mentioned in the ODD, although only in the 'stochasticity' section, not in the 'submodel' section (www.comses.net, a). According to the creators of the ODD, the 'submodel' section should include details like this (Grimm *et al*. 2006, 119).

Thirdly, a change was made to the calculation of maximum stock size of traders. In the ODD, this procedure is described as follows: "*For each trader the maximum-stock-size dependent variable is calculated as the average of the demand of the* other *traders he knows commercial information of, minus his own demand, rounded*" (www.comses.net, a; emphasis mine). In the original code, it is equal to the average demand minus the trader's demand, rounded, as one would expect. However, the calculation of average demand includes the demand of the trader itself. In other words, not just the average demand of "the other traders" is used to calculate the maximum stock size, but also of the trader in question itself. The replication was changed to conform with the original code.

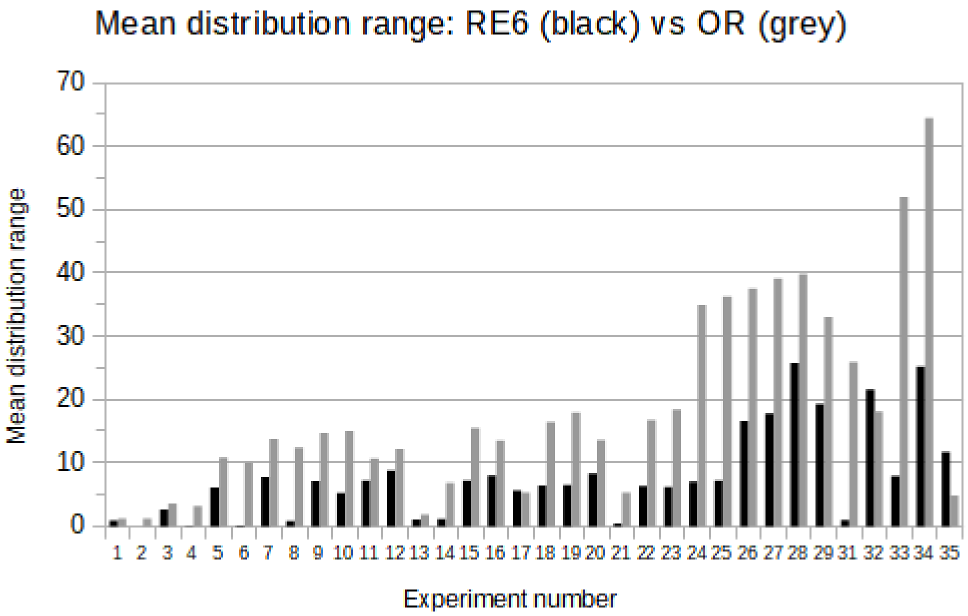

Figure 13: A bar graph showing the mean distribution range of all 100 iterations per experiment from version six of the replication next to the same measure from the original model, sorted by experiment number.

The data from version 6 of the replication can be found in appendix 13. As can be seen in figures 12 and 13, the ware distribution width and range of the replication is still consistently much lower than the original's. However, the percentage change of the mean distribution width has improved. On average they are -75,71% in version six, compared to the -80,73% of version 5. In the previous versions of the replication, in the first twelve experiments, it showed that experiments with higher *local-knowledge* value had a much lower average distribution width and range than their counterparts. In the current version the difference between these pairs of experiments has grown, but, it has grown because the average distribution width of the experiments with a *local-knowledge* of 0,1 has increased. This difference can be seen by comparing figure 12 and 10. However, this change makes the internal differences between the experiments bigger. What I mean by this is that the pairs of experiments, number one to twelve, with alternating *local-knowledge* values of 0,1 and 1 now have greater differences between them. However, the one's with a low *local-knowledge* value are closer to the original. Nevertheless, all average distribution width numbers are still extremely off. With a few exceptions, namely experiments 17, 32 and 35, the average distribution range also shows a much lower value in the replication.

At this point, it is clear that there is still something fundamentally wrong with the replication. The code that dictates production of wares and setting of prices has already been checked, so the trading of wares will be looked at for the next version.

3.7 Version 7

In version 7 of the replication one very small change, with major consequences, was made to the trading procedures. The code for version 7 of the replication can be found in appendix 14.

During each loop, every piece of tableware is put up for trade. This process includes a specific line that checks whether they can break even on the transaction. If so, the item is sold, and if not, the item is stored so that his process can be repeated during the next time step. This line includes an if-statement on whether the buyer's estimated price is equal to or greater than the seller's estimated price. I made the crucial mistake of reversing this operator, so that the trader checks whether the buyer's price estimation is less than their own instead. This mistake was then repeated for the trade procedure of each tableware type; the code was copied and only the tableware type was changed. Note that the comment after the code does state that it should check whether the buyer's price is greater than or equal to the seller's and the reverse, which means that the error was merely a very unfortunate typo and not a misunderstanding of the model.

The data form version 7 of the replication can be found in appendix 15. As can be seen in figures 14 and 15, both mean distribution width and range is much closer to the original in version seven. In the previous version, the median percentage change of average distribution width was -85,46% and in version 7 it is 0,71%. Distribution range also shows a similar huge change. In terms of average distribution width, many experiments are now within 2% percentage change in either direction compared to the original. There are also experiments with a greater difference in distribution, especially in the later experiments, with the highest percentage change

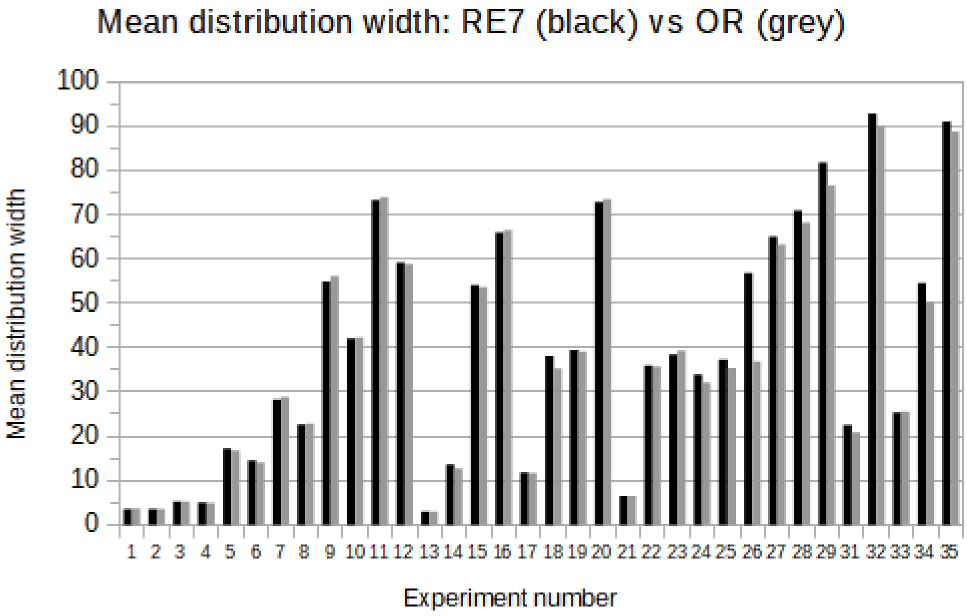

Figure 14: A bar graph showing the mean distribution width of all 100 iterations per experiment from version seven of the replication next to the same measure from the original model, sorted by experiment number.

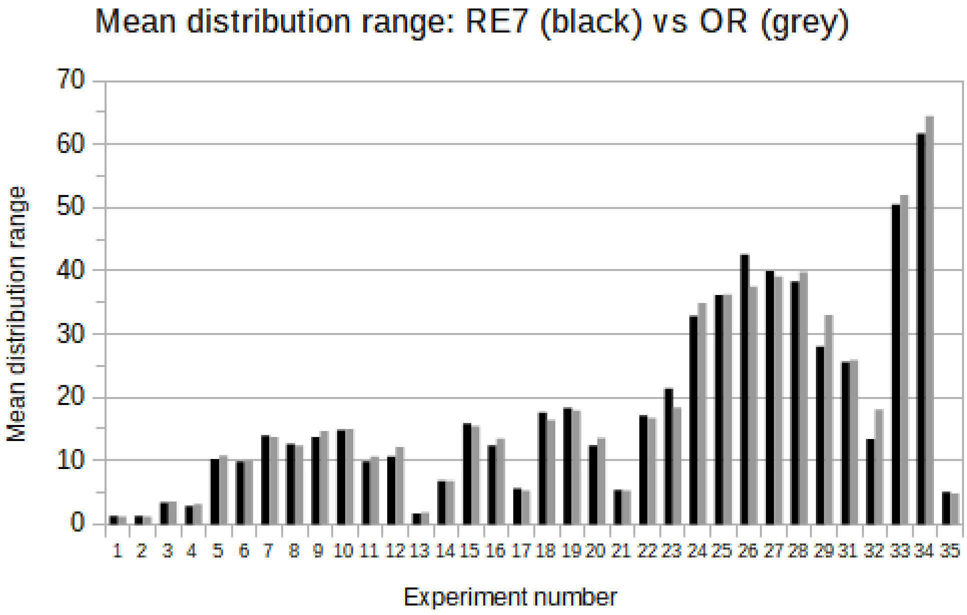

Figure 15: A bar graph showing the mean distribution range of all 100 iterations per experiment from version seven of the replication next to the same measure from the original model, sorted by experiment number.

being 8,33%. Experiment 26 is a major exception. It has a much higher average distribution in the replication: a percentage change of 53,78%. This is kind of odd, as experiment 26 does not use any extraordinary independent variable values.

This massive improvement between version 6 and 7 is, of course, very understandable, knowing the fundamental error that was fixed in this version of the replication. The minima, maxima, and mode still have big differences compared to the original in many cases, but I think this is inherent to these measures, especially the mode, which is quite useless when almost all data points in an experiment are different.

When analysing at the data from version 7, a small problem was noticed in the network measure data. The network measures of certain experiments that should be the same, i.e. experiments that had the same independent variables that govern network creation, had very small differences between them. This was not noted before, because I was comparing the data from the replication to the original, not data from experiments of the replication to each other. Before using more rigorous statistical methods to compare the original to the replication, it is my aim to fix this problem.

3.8 Version 8

Version 8 of the replication includes a fix to the aforementioned problem, an inconsequential change to the *traders-production-site* = '30,1,1,1' experiments, as well as some cleaning up of the code. The code for version 8 can be found in appendix 16.

Because the experiments that share the same independent variables that determine network creation also share the same random seeds, the only possible explanation for a difference in the network measures between these experiments is, to my knowledge, a external source that uses a different method of generating random numbers. Looking back at the data from the previous versions of the replication, I noticed that this problem started with version 5. In this version, the new and fixed method of connecting clusters was introduced, which uses an imported *WeakComponentClusterer* algorithm from JUNG. This algorithm uses Java, a programming language I am not familiar with, and so I decided to contact the Repast mailing list for help. I was advised to reimplement my own version of the algorithm that uses *LinkedHashSet* instead of *HashSet* (sourceforge.net). As I understand it, the *WeakComponentClusterer* algorithm uses a *HashSet*, a list of elements, as its random number generator instead of a seed. When using *HashSet*, this list is iterated unpredictably, but when using *LinkedHashSet* the iteration is always in the same order. I downloaded the source code, made the suggested change and added the class to my model. This fixed the problem at hand.

In addition to this change a minor change was made to the code that determines the assigning of 30 traders to one site and only one trader to the others. In one of the articles by Brughmans and Poblome (2016a), it is mentioned that production site A should always be the one to receive 30 traders. Before version 8, the site that was assigned more traders was chosen randomly in the replication. However, this should not make a difference, because when exporting the data, wares types are ordered by the amount of sites they appear on the type labels are disregarded entirely.

The data from version 6 of the replication can be found in appendix 17. Since the network creation process was altered slightly, the network measures should be discussed again. Although there are differences between the network measures of version 8 and those of versions 5 to 7, these differences are very minimal, as the connecting of components only results in the creation of a few links. In terms of the orignal's network measures falling between the minimum and maximum of the replication's, as was tested for version five, the same exceptions exist in version 8 as they do in versions 5 to 7; the clustering coefficient of experiment 34 and the average shortest path length of experiment 21 of seed 10 of the original model fall outside the ranges produced by the replicated model. Since the mean distribution width and range has changed so little between version seven and eight, no additional bar graphs will be presented. Instead, statistical tests will be employed as a more rigorous form of comparison with the original model. For the sake of structure, these tests will be discussed in the following section.

3.9 Statistical comparison

As mentioned in the first chapter, a model can be matched on several levels: numerical, relational and distributional (Axtell *et al*. 1996, 135). In this thesis, distributional equivalence is aimed for. Distributional equivalence can be shown by performing statistical tests. Following Axtell *et al*. (1996) and Miodownik *et al*. (2010) the Mann-Whitney U test is used compare the distributions of the data from the original and the replication in order to determine whether this data can be said to be distributionally equivalent. Because not all the data is normally distributed, t-tests can not be used. The null-hypothesis of the Mann-Whitney U test is the equivalence of the distributions of both populations. Note that the point of this study is to replicate the original model, and thus the aim is not to reject the null-hypothesis of the Mann-Whitney U test, but to accept it. Two sets of Mann-Whitney U tests were performed to compare the tableware distribution values of the 34 experiments: one of the average widths of all four tableware types (tab. 5) and one of the range as defined by Brughmans and Poblome (2016b), i.e. the maximum width across all tableware types minus the minimum width (tab. 6). The width and ranges were chosen because both of these statistics were used by Brughmans and Poblome (2016a) to compare the experiments to the pattern found in the ICRATES data. The average of all product types was tested as a whole instead of performing four different tests for the four tableware types. This reduces the amount of tests that have to be analysed while still addressing the point of the width statistic, namely, the number of products that were dispersed. All statistical calculations were performed using jamovi, version 0.9.1.12 (www.jamovi.org).

The results of the Mann-Whitney U test of the average width can be found in table 5. In eight cases, the p-value is higher than 0,05, and thus the null-hypothesis cannot be rejected for the majority of the experiments. The exceptions are experiments 6, 14, 18, 26, 29, 32, 34 and 35. The Mann-Whitney U test of the range values, shows a greater equivalence between the replication and the original. The null-hypothesis also cannot be rejected for the majority of the experiments. In this case, the exceptions are experiments 23, 29, 32 and 34.

Table 5: The results of Mann-Whitney U tests of tableware average width from the original model versus version 8 of the replication (n=100 for every test).

Exp.	Mean OR	Mean RE8	statistic	*p*
1	3,67	3,60	4271	0,068
2	3,57	3,59	4875	0,755
3	5,29	5,33	4731	0,509
4	4,95	5,01	4831	0,678
5	16,9	17,2	4504	0,225
6	14,1	14,7	4169	0,042
7	28,8	28,4	4877	0,764
8	22,9	22,7	4848	0,710
9	56,1	54,9	4363	0,120
10	42,3	42,2	4887	0,783
11	74,0	73,2	4399	0,142
12	58,8	59,3	4700	0,464
13	2,95	2,94	4954	0,909
14	12,7	13,7	3608	<,001
15	53,7	54,0	4825	0,670
16	66,5	66,3	4645	0,386
17	11,7	11,8	4833	0,683
18	35,4	38,1	3805	0,003
19	39,2	39,5	4845	0,705
20	73,6	72,9	4491	0,214
21	6,51	6,54	4798	0,622
22	35,8	36,0	4711	0,481
23	39,5	38,4	4259	0,070
24	32,2	34,1	4556	0,278
25	35,5	37,5	4536	0,257
26	37,0	56,8	1184	<,001
27	63,3	65,1	4710	0,479
28	68,2	71,1	4505	0,226
29	76,6	81,8	3793	0,003
31	20,9	22,6	4340	0,107
32	90,1	92,9	4069	0,023
33	25,6	25,4	4688	0,446
34	50,4	54,0	2043	<,001
35	88,8	91,1	2474	<,001

Table 6: The results of Mann-Whitney U tests of tableware range values from the original model versus version 8 of the replication (n=100 for every test).

Exp.	Mean OR	Mean RE8	statistic	*p*
1	1,26	1,33	4889	0,752
2	1,21	1,29	4572	0,231
3	3,58	3,43	4579	0,294
4	3,18	2,97	4500	0,210
5	10,8	10,2	4559	0,280
6	9,98	9,61	4673	0,423
7	13,7	13,8	4934	0,873
8	12,4	12,5	4883	0,775
9	14,7	13,5	4561	0,283
10	15,0	14,6	4917	0,839
11	10,6	9,79	4553	0,274
12	12,2	10,9	4334	0,103
13	1,83	1,72	4679	0,404
14	6,83	6,77	4894	0,796
15	15,5	15,8	4789	0,607
16	13,5	12,3	4621	0,354
17	5,33	5,64	4559	0,277
18	16,4	18,0	4301	0,087
19	17,9	18,3	4899	0,805
20	13,6	12,5	4694	0,454
21	5,37	5,32	4884	0,776
22	16,7	16,9	4851	0,715
23	18,4	21,4	3981	0,013
24	34,9	33,4	4784	0,598
25	36,3	35,7	4757	0,553
26	37,6	42,9	4236	0,062
27	39,1	39,6	4956	0,914
28	39,8	38,0	4710	0,478
29	33,0	27,9	4060	0,022
31	25,9	25,9	4974	0,950
32	18,1	13,4	3860	0,005
33	52,0	49,5	4272	0,075
34	64,5	62,4	3793	0,003
35	4,88	5,15	4796	0,615

Table 7: The results of Mann-Whitney U tests of tableware average width from random seeds 1-50 versus random seeds 51-100 of the original model by Brughmans and Poblome (2016a) (n=50 for every test). The raw tableware data was shared by Tom Brughmans via email (appendix 1).

Exp.	Mean 1-50	Mean 51-100	statistic	*p*
1	3,66	3,69	1175	0,599
2	3,56	3,59	1145	0,459
3	5,18	5,38	1161	0,539
4	4,82	5,09	999	0,081
5	16,5	17,2	1089	0,268
6	14,0	14,1	1216	0,814
7	28,8	28,9	1203	0,746
8	22,4	23,5	935	0,030
9	55,8	56,5	1124	0,387
10	42,4	42,2	1232	0,904
11	74,3	73,8	1177	0,617
12	58,9	58,8	1226	0,869
13	2,92	2,98	1142	0,454
14	12,5	12,9	1110	0,335
15	53,3	54,1	1150	0,490
16	65,4	67,6	924	0,025
17	11,6	11,8	1141	0,454
18	36,9	33,8	927	0,026
19	38,9	39,5	1181	0,637
20	72,2	75,0	843	0,005
21	6,58	6,45	1220	0,839
22	35,5	36,2	1105	0,317
23	39,6	39,3	1185	0,654
24	32,5	31,9	1181	0,637
25	35,8	35,1	1194	0,699
26	37,9	36,0	1107	0,324
27	64,6	61,9	1112	0,343
28	68,4	68,0	1206	0,764
29	77,2	75,9	1152	0,499
31	21,2	20,6	1131	0,414
32	91,0	89,2	1080	0,242
33	25,3	25,9	1028	0,125
34	50,1	50,6	1114	0,350
35	88,5	89,1	1116	0,355

Table 8: The results of Mann-Whitney U tests of tableware range values from random seeds 1-50 versus random seeds 51-100 of the original model by Brughmans and Poblome (2016a) (n=50 for every test). The raw tableware data was shared by Tom Brughmans via email (appendix 1).

Exp.	Mean 1-50	Mean 51-100	statistic	*p*
1	1,22	1,30	1160	0,478
2	1,12	1,30	1084	0,186
3	3,68	3,48	1154	0,502
4	3,08	3,28	1147	0,468
5	10,5	11,1	1176	0,609
6	10,1	9,90	1231	0,898
7	13,1	14,4	1079	0,237
8	12,2	12,7	1192	0,691
9	14,1	15,2	1141	0,452
10	14,2	15,9	1079	0,239
11	9,92	11,3	1059	0,188
12	12,2	12,2	1182	0,641
13	1,78	1,88	1164	0,528
14	6,38	7,28	1024	0,118
15	16,0	14,9	1154	0,507
16	13,5	13,4	1214	0,806
17	5,08	5,58	1058	0,181
18	16,6	16,2	1179	0,624
19	17,4	18,5	1139	0,446
20	13,2	13,9	1152	0,499
21	5,24	5,50	1105	0,314
22	16,4	17,0	1154	0,507
23	18,5	18,3	1237	0,931
24	34,5	35,4	1199	0,728
25	35,9	36,7	1232	0,904
26	36,0	39,2	1079	0,238
27	37,1	41,2	1033	0,135
28	38,1	41,6	1109	0,333
29	30,8	35,3	1071	0,217
31	25,5	26,2	1207	0,769
32	15,5	20,6	1030	0,129
33	52,0	52,1	1246	0,981
34	65,1	63,9	1140	0,447
35	4,80	4,96	1212	0,791

After finding that the distributions of the range and average width cannot be said to be drawn from the same population in all cases, the code of the replication was checked and compared to the original several times again. However, no more differences were found that could have any impact on the results of the model. This raises the question: what causes the discrepancy between the original model and the final version of the replication?

Before attempting to answer this question, I wanted to test whether the Mann-Whitney U tests are a proper way of testing the difference between the replication and the original. A similar set of tests as above was performed, but this time only data from the original model was used. This data was grouped by random seeds 1 to 50 and 51 to 100. The pairs of experiments of the average distribution width show that four experiments, 8, 16, 18 and 20, appear to be drawn from statistically different distributions (tab. 7). This suggests that average distribution width might not be an adequate variable for testing whether the replication is statistically equivalent to the original. This is not the case for the Mann-Whitney U tests of range values; the null-hypothesis cannot be rejected in any case (tab. 8).

3.10 Discussion

Returning to the issue of what could cause the discrepancy between the original model and the final version of the replication, one possibility is that there are differences between the original and the replication that were not noticed when the models were compared. I do not find this explanation satisfactory, as the model was checked multiple times throughout the replication process and again after the statistical tests of replication version 8. Another possibility is that original model whose output data was used in the papers by Brughmans and Poblome (2016a; 2016b) are not the same as the model from which the source code and ODD was published (www.comses.net, a). It can be said for a fact that there is at least one difference between the two because in supplement 1 of the JASSS paper (Brughmans and Poblome 2016a), a *transport-cost* variable is mentioned, which was only used in experiment 30. This experiment was not discussed in the paper and no such variable can be found in the ODD or the source code. In a correspondence with Brughmans (appendix 19), he told me that he did some experimenting with this variable for publication elsewhere but that this version of the model was not yet documented. I therefore assume that the model published online is otherwise identical to the one discussed in both papers and that the *transport-cost* entry in the supplement table was an artefact of this test version. A third possibility is that the discrepancies in output data are due to inherent differences between the NetLogo and Repast Simphony ABM toolkits. Bajracharya and Duboz (2013) modelled a simple epidemiological model in three ABM toolkits, NetLogo, Repast and Cormas, and compared the results. They found that "*the agent-based platforms did not give the similar simulation results for the same model with the same set of experiments*", stating differences in scheduling, list sorting and shuffling mechanisms as possible causes (Bajracharya and Duboz 2013, 5). With this in mind, it is interesting to note that experiments with random network structures appear to be overrepresented among the experiments that have a statistically significant distribution difference between the original and the replication. In the sets of tests of the range values, three of the four experiments with a statistically significant distribution difference are experiments with random network structures. In the set of tests of the width averages, this number is four out of eight. Note that experiments with random network structures only make up four out of the 34 experiments. Were the differences between the output data of the replication and the original model caused by differences in how the toolkits that were used handles the random selection of agents from a list? Attempting to answer this question goes beyond the scope of this thesis, but it should certainly be considered as a possibility.

Experiment 26 is still an anomaly: it has a percentage change in average distribution width from the original to version 8 of the replication of 53,82%, which is far greater than the second highest percentage change of all the other experiments. Since this experiment does not use any extraordinary independent variable values (see tab. 3), I do not believe that this deviation can be explained by differing ABM toolkits. It is possible that the independent variables used for experiment 26 do not match with the original.

In summary, the replication process has shown several causes of discrepancies between the various versions of the replicated model and the original. The ODD was imprecise in several regards. A recurring issue was the timing of checking requirements for link creation between traders. The ODD was written in a manner that led me to believe that randomly selected pairs of traders should be selected if they meet certain requirements, but in the original model, pairs of traders that are to be linked are selected for using said requirements. This seemingly small distinction led to differences in the network of traders. In other cases, the ODD was either missing some details, details were included in the wrong section or different sections of the ODD contradicted each other. However, not all differences between the replication and the original were due to inaccuracies in the ODD. My own inexperience with coding, let to a bug in the trading procedures with major consequences for the ware distribution. However, in the end the replication was for the most part successful, as distributional equivalence was shown for the majority of experiments. My broader criticisms of the MERCURY model, as they were formed during the replication process, will be presented in the next chapter.

4 Critiques of MERCURY

A chapter on critiques of MERCURY might seem unrelated to to the main topic of replication of the model, but checking model verification, which many of these critiques fall under, is an important part of replication in agent-based modelling (Wilensky and Rand 2007). Besides, forming my critiques was only possible by engaging deeply with MERCURY, through the process of replication. Before discussing my own critiques, I would like address an existing response to the MERCURY model by Van Oyen (2017) and a subsequent reply by the original creators of the model, Brughmans and Poblome (2017).

4.1 Existing critiques of MERCURY

Van Oyen (2017, 1356-1357) describes agent-based modelling, with its need for rigidly defined variables, as being fundamentally modernist. In addition, she states that the MERCURY model in particular contains modernist elements, such as profit-seeking behaviour, the lack of a spatial dimension of 'markets' and a divide between the social, i.e. the sharing of information between traders, and the economic, i.e. the ability to trade (Van Oyen 2017, 1357-1358). She questions whether a model such as Bang's (2008), which she describes as primitivist, can ever be found to be more likely to be correct than a modernist one, such as Temin's (2012), using these inherently modernist methods. Van Oyen (2017, 1358) proposed several potential changes to the model, which might even out the playing field, namely: an emphasis on social bonds over profit, creating additional dependencies between variables in order to lessen to divide between the social and the economic, and adding a spatial dimension to 'markets'. She goes on to say that different objects, tableware in this case, behave according to different logic, taking issue with the universalising of commodities. This qualitative difference between objects is not addressed by the quantitative nature of Brughmans and Poblome's (2016a; 2016b) research (Van Oyen 2017, 1359-1360). Lastly, she discusses whether the emergence of different, specialised, trader agents is a possible avenue for MERCURY, such as is observed in Roman agricultural storage facilities (Van Oyen 2017, 1360-1361).

In a short response to Van Oyen, Brughmans and Poblome (2017) state that they believe agent-based modelling can indeed result in support for primitivist hypotheses. They explain that agent-based modelling was developed to address past arguments against equation-based modelling, such as the absence of heterogeneity of modelled objects and assumptions of global knowledge and profit maximisation, arguments which Van Oyen makes in regard to agent-based modelling. Brughmans and Poblome agree with Van Oyen's arguments regarding the need for multiple agent types and the importance of debating agent-based modelling and its assumptions.

Although I agree with Brughmans and Poblome's (2017) argument that agent-based modelling does not have inherently modernist assumptions and that it can be used to support primitivist hypotheses, the code of MERCURY does explicitly contain the modernist assumption of profit maximisation. As discussed in the previous two chapters, whenever agents trade, they mathematically compare whether the buyer's estimated price of a product is equal to or higher than the seller's; if it is they trade and if it's not the seller stocks the product for the next loop (www.comses.net, a). This is also specifically mentioned by Van Oyen (2017, 1357), but not addressed by Brughmans and Poblome's (2017). However, I do not share Van Oyen's concern that support for Bang's (2008) Roman bazaar model might not be able to arise from the MERCURY model because of this, and other, modernist assumptions. I believe that Bang's bazaar model should not be seen in such black-and-white terms as 'primitivist' or 'modernist'. Bang (2008, 28-29) himself has issues with how the debate on ancient economies is so restricted to the primitivist and modernist concepts. He does not disagree with the importance of the market in the Roman empire, with which I assume him to imply the existence of a profit maximising mindset among traders, but he rejects the notion that the Roman economy was comparable to modern market economies in terms of the availability of information and network integration (Bang 2008, 4-5; 35; 137-139; 295). Therefore, I believe the assumption of profit maximisation in MERCURY is not at odds with Bang's bazaar model and that Brughmans and Poblome's (2016a; 2016b) representation of Bang's model is valid in this regard. Van Oyen's other criticisms of the MERCURY model, namely the lack of a spatial dimension and the point regarding multiplicity of agents, are, in my view, valid. It should be noted that the lack of a spatial dimension in MERCURY's network is likely to be addressed in a future publication, as an extension, which includes a *transport-cost* variable that could remedy this issue, was published online (www.comses.net, b).

4.2 New critiques of Brughmans and Poblome's research

This section contains my own critiques of Brughmans and Poblome's (2016a; 2016b) research, concerning both the MERCURY model itself and how it reflects the differences between Bang (2008) and Temin's (2012) models, as well as the conclusions that are drawn from the data. I will also suggest ways in which MERCURY can be modified in the future.

Firstly, I would like to address the role that commercial information plays in the model. The availability of commercial information is represented by the *local-knowledge* variable, which determines the proportion of other traders one trader is connected to from which they know their *demand* and supply (Brughmans and Poblome 2016a). However, this information is only used in setting a trader's price estimate and maximum stock size. It is not used for determining how much tableware is produced each time step. When producing new tableware, traders on production sites check if their total amount of products combined is less than their own *demand*, and if it is, they produce to meet their *demand*. This process is solely based on the trader's own demand and does not take into account the demand of other traders from which it has obtained commercial information. Why the use of commercial information available to traders is limited in this way is not explained by the authors. It would be logical for traders to use information about demand amongst their peers to determine how much production is needed. It is possible that the fact that the average demand amongst informants is not used to determine production was intended to be a reflection of the poverty of information across regional boundaries, but this is not discussed by the authors. As a minor side-note, Brughmans and Poblome (2016a) describe that with a *local-knowledge* value of 0,1, traders receive commercial information from 10% of their neighbours. In practice, this is not accurate. The average amount of links each trader has is roughly 4,5, and because the amount of neighbours a trader receives information from is rounded up, to a minimum of one, the average amount of informants each trader has is, in actuality, closer to 22%.

Secondly, the way in which the creation of the network is limited to a maximum number of links, misrepresents the differences in integration between markets in Bang (2008) and Temin's (2012) models. When the network is being formed, first inter-site links are created and afterwards, links are created between traders on the same site until a certain average number of links per trader, the average degree, is reached. In the experiments performed by Brughmans and Poblome (2016a), the average degree is constant throughout all of them. This means that, regardless of the independent variables that are used, each experiment has approximately the same degree of connections between traders. The integration between markets differs depending on how many inter-site links are created. However, because the amounts of links that are created is limited, increasing the amount of inter-site links decreases the amount of intra-site links on the same site, as acknowledged by Brughmans and Poblome (2016a). Note that, for the sake of this argument, I count links between mutual neighbours as 'intra-site links' because they are also links between traders on the same site. If the amount of inter-site links that will be created is higher, representing Temin's model, inter-site integration will be higher, but intra-site integration will be lower. While it is true that in Temin's (2012) model market integration across the Roman empire is high, at least compared to Bang's (2008) model, this does not necessarily mean integration was low on a local scale, and, as far as I can tell, this is not suggested by Temin (2012). By representing Temin's model in this way, this model is favoured, because traders have a much higher chance to sell tableware to buyers on other sites as the ratio of potential buyers is skewed higher towards traders on other sites than traders on the same site as the seller. This would not be the case if intra-site links are created first, until a certain average degree is reached, before connecting traders on different sites.

Thirdly, the range calculations do not tell the whole story. In Brughmans and Poblome's (2016a; 2016b) papers, the range of distribution is calculated as the number of sites on which the ware with the widest distribution is deposited minus the number of sites on which the ware with the lowest distribution width is deposited. Brughmans and Poblome (2016b, 403) use this calculation of range values to claim Temin's model is closer to the archaeologically observed pattern than Bang's model: "*Only high proportions of inter-site links, representing a high integration of markets (as argued by Temin 2013), have the potential to give rise to the archaeologically observed differences in the width of tableware distributions.*" However, if instead the percentage difference between the pair of sites with the highest and lowest distribution width is used, this conclusion does not completely hold up. When using an absolute difference, you favour experiments with a higher overall width. Considering that the simulation and its output data cannot directly equate to the archaeological data, I believe looking at the percentage difference between tableware distribution width is at least as useful, if not more so, than simply looking at the absolute difference between the two. Another criticism one could make of Brughmans and Poblome's (2016a; 2016b) distribution range calculations is that it only takes into account the tableware types that are located on the most and least amount of sites. The pattern of interest is of "*one product being much more widely distributed than the three other products*" (Brughmans and Poblome 2016a). If only the least and most widely distributed tableware types are used to determine the range of distribution, this pattern could be falsely identified. An extreme example would be if two hypothetical experiments produced average product widths of 100, 95, 90, 10 and 100, 20, 15, 10. Using Brughmans and Poblome's (2016a; 2016b) range of distribution calculations, both ranges would be equal even though the distribution patterns are completely different, the latter one being much closer to the archaeologically observed one. Such extreme cases do not occur in the papers, and the average product width of all products is included in all tables for comparison with the desired pattern, but many, if not most, experiments do show a gradual decrease in width from the most to least widely distributed wares, instead of the desired high

Table 9: A comparison of range values of experiments 1 to 12 as calculated by Brughmans and Poblome (2016a), by percentage difference between the most and least widely distributed product, the most minus the second most widely distributed product and the percentage difference between the most and second most widely distributed product.

Exp.	*local-knowledge*	*proportion-inter-site-links*	Range: original	Range: percentage difference	Range: 1st-2nd	Range: perc. diff. 1st & 2nd
1	0,1	0	1,26	33,12%	0,49	11,91%
2	1	0	1,21	32,38%	0,53	13,21%
3	0,1	0,0001	3,58	64,14%	1,58	23,78%
4	1	0,0001	3,18	60,79%	1,35	21,73%
5	0,1	0,0006	10,81	64,20%	3,96	19,18%
6	1	0,0006	9,98	70,83%	3,41	19,19%
7	0,1	0,001	13,72	48,68%	4,82	14,78%
8	1	0,001	12,42	54,35%	4,56	16,94%
9	0,1	0,002	14,68	26,78%	4,36	7,17%
10	1	0,002	15,02	36,44%	4,98	10,50%
11	0,1	0,003	10,62	14,42%	3,88	5,00%
12	1	0,003	12,16	21,03%	3,80	6,06%

difference between the most widely distributed wares and all the others. If the desired pattern is that one tableware type is much more widely distributed than the rest, comparing the most widely and second most widely distributed wares would be a better way to identify said pattern in the simulated data.

Tables 9 and 10 contain the results of experiments on which Brughmans and Poblome (2016a; 2016b) base their preference of Temin's model over Bang's, in regards to distribution range. Three extra measures of range were added: the percentage difference between the distribution width of the most and least widely distributed wares, the distribution width of the most minus the second most widely distributed wares and the percentage difference between the distribution width of the most and secod most widely distributed wares. Percentage difference is calculated as follows:

$$\frac{|width_a - width_b|}{\frac{width_a+width_b}{2}}\times 100\%$$

In the first table, pairs of experiments with varying *local-knowledge* and *proportion-inter-site-links* values are compared. When looking at the range values as defined by Brughmans and Poblome and the absolute range between the most and second most widely distributed wares, the second highest *proportion-inter-site-links* value, 0,002, produces the highest tableware ranges. If the percentage difference is used, a much lower value, 0,0001 and 0,0006, produces the highest difference, depending on which tableware types are compared. The second table contains the results of experiments with differing proportion *proportion-inter-site-links* and *maximum-demand* values. These experiments all share the fact that the number of traders on production sites is not fixed. Again, these experiments show that increasing *proportion-inter-site-links* results in greater range differences using the original calculation. However, all other methods of calculation range show that experiment 24, an experiment with a relatively low *proportion-inter-site-links* compared to the other experiments, has the highest distribution range. This contradicts Brughmans and Poblome's (2016b, 405) findings about the importance of market integration.

Fourthly, the limited variation between experiments performed weakens Brughmans and Poblome's (2016a; 2016b) conclusions. As previously discussed, the only input settings that produce results similar to the pattern perceived in the archaeological data is when the amount of traders on production sites is set to 30 traders on one site and only one on the remaining three. In Brughmans and Poblome's (2016a) words: "*The pattern observed in the archaeological data (i.e. that one product is significantly more widely distributed and the difference in distribution width between this product and the least widely distributed product (range) is high) was only reproduced in scenarios where one production centre has far more traders than any other production centre and the number of inter-site links is high (proportion-inter-site-links 0.001) (see experiments 33 and 34).*" The problem with this statement is that a highly unequal trader distribution among production sites was *only* tested in combination with a high *proportion-inter-site-link* value and with a randomly created network, which also results in a high amount of inter-site links. To properly make conclusions regarding the validity of Bang (2008) and Temin's (2012) models using MERCURY, varying *proportion-inter-site-link* values should be tried in combination with the aforementioned trader distribution. This is what I aim to do in the next section of this chapter.

Lastly, the scenarios of highly unequal trader distributions among production

Table 10: A comparison of range values of experiments 24 to 28 and 33 as calculated by Brughmans and Poblome (2016a), by percentage difference between the most and least widely distributed product, the most minus the second most widely distributed product and the percentage difference between the most and second most widely distributed product.

Exp.	*proportion-inter-site-links*	*traders-production-site*	*maximum-demand*	Range: original	Range: percentage difference	Range: 1st-2nd	Range: perc. diff. 1st & 2nd
24	0,001	na	10	34,92	104,39%	16,05	38,75%
25	0,001	na	30	36,31	100,81%	15,39	34,67%
26	0,002	na	10	37,59	102,57%	15,42	33,97%
27	0,002	na	30	39,15	67,58%	11,59	16,58%
28	0,003	na	10	39,84	64,05%	11,38	15,30%
33	0,001	(30,1,1,1)	10	52,05	142,81%	46,20	117,46%

Table 11: Average product width and ranges of the new experiments 36 to 40 and old experiment 33 for comparison. In these experiments, *proportion-inter-site-links* is varied. All other independent variables are kept constant: *equal-traders-production-site* = 'false', *traders-distribution* = 'exponential', *network-structure* = 'hypothesis', *local-knowledge* = 0,5, *traders-production-site* = '30,1,1,1', *max-demand* = 10.

Exp.	*proportion-inter-site-links*	Average product 1	Average product 2	Average product 3	Average product 4	Range: original	Range: perc. diff.	Range: 1st-2nd	Range: perc. diff. 1st & 2nd
36	0	5,07	4,52	3,99	3,32	1,75	42,30%	0,55	11,49%
37	0,0001	10,36	6,16	4,94	4,10	6,26	83,41%	4,20	47,73%
38	0,0006	41,82	12,57	9,72	7,33	34,49	139,65%	29,25	106,94%
33	0,001	60,35	16,69	13,67	10,83	49,52	138,81%	43,66	113,11%
39	0,002	84,73	25,85	22,02	17,99	66,74	130,36%	58,88	106,80%
40	0,003	92,52	33,55	29,20	24,63	67,89	116,43%	58,97	93,84%

sites bring with them problems of differences in production quantity. In MERCURY, during each time step, all traders on production site 'produce' tableware of the type corresponding to the production site they are on if the amount of products they own is less than their individual *demand* value. If one production site has thirty traders on it, while the others only have one trader on them, naturally the former production site's tableware type will be produced significantly more often. This is shortly acknowledged by Brughmans and Poblome (Brughmans and Poblome 2016a;2016b, 404), but I believe it deserves more attention. When one product type is produced much more often, it has the potential to flood across the network quicker than the other product types. While this might be an accurate reflection of the historical reality, you are not investigating the influence of network structure on tableware distributions at this point, but rather the differences in production quantity. This could be counteracted by altering MERCURY in a way that limits the amount of product that's produced, for example, by setting a limit to the amount of traders on a production site that can produce each time step, perhaps in combination with making the difference in the amount of traders on production sites slightly less extreme. It should also be mentioned that the precise numbers used, 30 traders to one production site and one to the others, are not justified by (Brughmans and Poblome 2016a). No reason is given for why this extreme difference would be more reflective of the archaeological reality than the exponential distribution to all sites, including production sites, that is used in other experiments.

4.3 Additional experiments and their results

In the previous section, I make the argument that by not performing experiments using varied *proportion-inter-site-link* values in combination with a highly unequal trader distribution among production sites, Brughmans and Poblome's (2016a; 2016b) conclusion that their research supports Temin's (2012) model over Bang's (2008) is less strong. Therefore, new experiments using these values were performed. These experiments were performed with the replicated version of MERCURY, because the '30,1,1,1' option is not available in the published version of MERCURY (www.comses.net, a). For these experiments, both the original way of calculating the range of distribution is used, as well as by percentage difference between least and most widely distributed products, the most widely distributed product minus the second most widely distributed product and the percentage difference between the most widely distributed product and the second most widely distributed product (tab. 11). The complete output data of these experiments can be found in appendix 18. No changes to the code of MERCURY were made, only the input variables were altered, so most criticisms discussed in the previous section still apply to these experiments.

The data from the new experiments conforms to Brughmans and Poblome's (2016a; 2016b) conclusions in part: increasing integration between sites results in more widely distributed products. In this regard, these experiments show further support for Temin's (2012) hypothesis. The original range calculation and the range calculation of the most widely minus the second most widely distributed product also suggest the experiments with the two highest *proportion-inter-site-links* values are closer to the archaeologically observed pattern, although the difference between these two experiments is marginal. However, similar to the results in tables 9 and 10, when assessing the range of distributions as a percentage difference, ranges peak at *proportion-inter-site-links* values of 0,0006 and 0,001 (tab. 11, experiments 38 and 33), depending on whether the most and least widely or most and second most widely distributed wares are compared. Regardless of the original or the new ways of calculating range of distribution is used, extremely small degrees of market integration (tab. 11, experiments 36 and 37) produce very small distribution ranges. If range distributions peak at a certain degree of market integration, one could ask the question: which values of *proportion-inter-site-links* correspond to Bang's (2008) and Temin's (2012) models? It would be difficult, if not impossible, to map specific values to non-numerical conceptual models. One could still make the argument that these new experiments support Temin's (2012) hypothesis over Bang's (2008), especially in regards to overall distribution width, but when taking into account ranges of distribution, Brughmans and Poblome's (2016a) claim that "*the emphasis on limited market integration in Bang's model is highly unlikely*" becomes less clear-cut.

In summary, the MERCURY model and the papers based on it, while innovative and interesting, can be criticised from various angles. In the past, Van Oyen (2017) has criticised the model mainly on a theoretical ground, arguing that it has a modernist bias, and that its lack of a spatial dimension and variability of traders and traded objects contradicts the archaeological reality. My own critique is mainly methodological, focusing on the details of the model and the way the data is interpreted. These critiques include: the inconsistent use of information by traders, the way in which the network of traders is created fundamentally favouring one hypothesis over the other, the limited number of experiments that were performed using an input variable that most resemble the archaeological data, and that this input variable which leads to a close resemblance to the archaeological reality causes massive inequality in production which counteracts the analysis of the network structure. Another criticism, and arguably the strongest one, is that the distribution ranges calculated in the original study do not properly help us identify the pattern as it is found in the archaeological data. Brughmans and Poblome's (2016a; 2016b) conclusion that Temin's (2012) hypothesis is strongly favoured over Bang's (2008), using the data from the MERCURY model, does not totally hold up using newly presented ways of calculating the range of distribution and using new data created in additional experiments with the replicated version of MERCURY. I hope criticisms such as these can help improve the MERCURY model in the future.

5 Discussion

In this chapter, I want to discuss how the replication of MERCURY presented in this thesis compares with other published replication studies in terms of methodology and results. Unfortunately, many replication studies are quite brief. They often only present a short overview of the replication process, before going into the results of the replication. The various versions and problems encountered along the way are mostly not written about. As a result, only a selection of replication studies which yielded interesting comparisons will be discussed here.

The often cited study by Axtell *et al*. (1996) introduced the replication standards of numerical identity, distributional equivalence and relational alignment. However, even though these standards, which are used in many replication studies, were introduced in this study, its methodology is not one of a traditional replication study. The method used by Axtell *et al*. (1996) is 'model alignment', also called 'docking', which has been described as "*an alternative method of replication*" (Romanowska 2015b, 182). Instead of implementing a new version of the model that is to be replicated, two models that are similar in many ways are chosen, and one of them is altered to produce the same results as the other one. The aim of this method is to show if two models that concern phenomena can produce the same outcomes (Axtell *et al*. 1996, 124), and as such it does not entirely match the aims of other replication studies, including this one. Axtell *et al*. (1996, 135) chose to test for distributional equivalence, which was shown for eleven out of the twelve data sets they produced (Axtell *et al*. 1996, 128-131). The authors believe that this difference is caused by the fact that a specific procedure of the model that was altered, was left unchanged. Axtell *et al*. (1996) are not very clear on whether their study counts as a 'successful' docking attempt, considering not all data sets were found to be distributionally equivalent. Miodownik *et al*. (2010) also report that some of the data sets from their replication can not be matched statistically with the original, providing mixed support for distributional equivalence. In this way, these two studies are similar to my experience; the majority of experiments are distributionally equivalent to their original counterparts, but not all. The ABM community does not seem to have clear established definitions of when a replication is 'successful' as a whole.

In a study by Donkin *et al*. (2017), two replications were made in two different ABM platforms: NetLogo and Repast Simphony, using Java. This is not the only way in which this study differs from my own; Donkin *et al*. (2017) were missing a lot of information, only the description in the original paper (Potting *et al*. 2005) was available. As part of their replication study, an ODD was created based on their interpretation of the description in said paper. They concluded that all three models, the two replicated versions and the original model, could not be matched with each other on any of the three levels proposed by Axtell *et al*. (1996). It should not be a surprise that the differences between the original and the two replications were caused by a lack of information about the original model. However, the differences between the two replicated models are noteworthy, especially in the context of this thesis. Donkin *et al*. (2017, 150), concluded that this was likely caused by inherent differences between the programming languages; Java is a low-level language and thus requires more manual defining of processes. Note that Donkin *et al*. (2017) used the Java programming language instead of Groovy and ReLogo, which were used in this thesis, so this situation is not entirely equivalent. However, this does bolster the claim that different programming languages can produce significant differences in replicated models, as was already shown by Bajracharya and Duboz (2013), and might be the cause of the discrepancies between MERCURY and the final version of my replication.

In a conference report about their replication of an economic agent-based model, Legendi and Gulyas (2012) talk about how they isolated and verified specific parts of the model separately. I found myself doing the same thing, by first trying to match the network creation before moving on to the production and trade parts of MERCURY. I believe this is a useful technique, as it allows one to exclude parts of the model when seeking the issues in the code, which speeds up the replication process. This approach might not always be possible, but if it is, I recommend others to follow this procedure.

Edmonds and Hales (2002) published a replication study wherein each author created a replication of a model independent of each other in different programming languages, and compared the three resulting data sets, one from the original and two from the replications. Although no reference is made to the standards by Axtell *et al*. (1996), their methodology can be described as seeking distributional equivalence, i.e. statistical tests are used to determine whether the results can be said to be from the same distribution or not. Edmonds and Hales (2002), among other things, found that in the original model a certain unwanted bias existed when selecting agents that was not explained in the model's description. By implementing the model twice, they were easily able to identify that this was a fault of the original model instead of their replication. Edmonds and Hales' (2002) approach of a double replication, by separate authors, is very advantageous. I believe that if the MERCURY model were to be replicated twice, the possibility of discrepancies being caused by differences in the programming languages could be resolved, although this would not necessarily be dependent there being two separate model authors.

Small differences in the code producing noticeable disparities in the output data was not only apparent in Edmonds and Hales' (2002) study, but also in Wilensky and Rand's (2007) research. In this study, issues of timing, misrepresentations of the original model description and whether certain lists were shuffled or not caused statistically noticeable discrepancies between the original and the replication. Similar issues were found in my replication study, especially in the network creation process, as described in chapter three. This once again shows that very precise model descriptions are necessary, as even slight differences can result in substantive disparity between the replication and the original model. It also provides a greater insight into the hidden assumptions of the model. After a replication attempt, the original authors might go back and alter their model if they find that these hidden assumptions are not valid.

Statements regarding the importance of publishing detailed descriptions of agent-based models and sufficient output data to test for equivalence are ubiquitous in replication studies (Axtell *et al*. 1996, 135; Donkin *et al*. 2017; Edmonds and

Hales 2002; Wilensky and Rand 2007). Insufficient comparison data and model descriptions was also a slight issue in this study, although in general the ODD protocol (Grimm *et al*. 2006; Grimm *et al*. 2010) has remedied this problem to a major extent. It should be noted that this issue also prevails in other computational subfields of archaeology. The fact that in his paper about replication of computational archaeological research, Marwick (2017, 445) points out that the publishing of data is relatively prevalent for ABM studies is very telling, in my view.

I want to end this chapter by providing suggestions to researchers who publish ABM studies and those who wish to engage in replication studies. It has already been mentioned, but I want to stress again the importance of publishing the data and resources required for replication. When choosing an agent-based model to replicate, I went through many papers that either did not include a link to their source code at all, or the link was dead. In fact, the papers about MERCURY were some of the only ones that included a working link to the source code of their model. CoMSES / OpenABM provides a service where anyone can upload their models free of charge, which I suggest researches to use. Perhaps it should be customary to include a secure hash, cryptographic code that can be used to determine if two files are the same, so readers can be sure the version of the model that was published is the same as the one that was used to write the paper. Output data too should be published in its entirety. If no data, or only summary statistics of the output data, is released a replication can never be confirmed to be equivalent on the distributional level, as this requires statistical comparison of the output data. Of course, Brughmans was kind enough to share this data (appendix 1), but ideally, it should be included as a supplement. Related to this is the explanation of the model in the ODD. The ODD was designed to facilitate replication, and as such it should include a very detailed explanation of the model's processes. If even some small details are missed by the person performing the replication, either because they were not included or because they were in the incorrect section of the ODD, this can result in great differences in the output of the model. As for advice for those who are replicating models, I will actually go against a standard that was discussed in chapter one, namely the requirement to use a different toolkit or programming languages when replicating. It is true that in an ideal situation, one should use a different ABM toolkit and/or language to perform a replication, as specific toolkits can unknowingly introduce bias for a conceptual model that is being modelled, which could be revealed by replication using a different toolkit. However, in practice, actually confirming that differences between the replication and the original data is a result of the difference in the toolkit that is used is very difficult, as this would require intricate knowledge of both toolkits. It might not even be possible if the software in question is not open-source. If the same toolkit is used, all the other ways in which replication is useful, such as confirming the accuracy of the ODD and aiding in model verification and validation, still apply; only one potential source of bias is ignored. Using the same toolkit also has an advantage: when using the same ABM toolkit, of the same version, and the same random seed values, barring any minor externalities, one can make sure that any differences in the output are due to a difference in the code alone, which might have its origin in inaccuracies in the ODD or bugs in the code, either in the original or the replication.

6 Conclusions

The specific purpose of this study was to replicate the MERCURY model in order to check whether the published description of the model is accurate and the results are not reliant on local conditions. The research questions, defined in the first chapter, will be addressed here.

The first research question was as follows: *Can an independent replication of the MERCURY model match the results presented by Brughmans and Poblome* (2016a) *on a distributional level, as defined by Axtell et al* (1996)*?* Distributional equivalence can be shown by statistically comparing the output data of the replication and the original. Both the average distribution width of all four product types (tab. 5) and the range of distribution (tab. 6) was tested using Mann-Whitney U tests. As shown in chapter three, section eight, distributional equivalence between the original and the final version of the replication can be shown for a majority of the experiments for both sets of statistical tests. Exceptions do exist; the difference in distribution is statistically significant for the average width in eight out of 34 experiments and in four out of 34 for the tests of distribution range. However, it should be noted that the former test might not be completely appropriate, as it also shows statistically significant differences within the same experiments of the original, when they are subdivided by random seed (tab. 7). There are no guidelines as to what proportion of replicated data sets have to statistically conform with the original in order for the replication as a whole to be labelled distributionally equivalent (Axtell *et al*. 1996), and as such I conclude that my data provides partial support distributional equivalence.

The second research question goes deeper into the specifics of replication process: *Can this replication be performed based solely on the description in the ODD, and if not, what are the shortcomings of the ODD?* Although the explanation in the

ODD was crucial in this replication attempt, there were definitely points where it fell short. The lack of specificity in the ODD caused several errors along the way. Some of these were only minor, but others caused greater dissimilarities. One way in which the ODD was unclear, was how the requirements for creating links worked. In multiple cases in the code, the requirements were used to select the traders which would be linked, instead of selecting traders and then checking if they meet the requirements for link creation. The details of this were not adequately explained in the ODD and a literal interpretation would suggest the latter way was used. In certain cases, specifics about submodels were not explained in the appropriate section. In another case, descriptions in the submodel section contradicted a description elsewhere; for the connecting mutual neighbours, in the submodel section it was mentioned that the initially selected traders were selected uniformly, while in the 'stochasticity' section the selection was explained as proportional to the amount of neighbours each trader has. Even though I had read the ODD in its entirety, while coding I mostly used the descriptions in the submodel section as a guide, as this section should include a detailed explanation of the submodels, which resulted in differences with the original model. Statistical tests were only made for the final version of the replication, as there existed a great error in the code that skewed the results too much for statistical tests to be meaningful, so the exact influence of each of these errors caused by inaccuracies in the ODD, or my interpretation of it, was not determined. However, some of these errors influenced network creation to a great amount, so I believe it is safe to assume that they would have also influenced ware distribution.

The third research question concerns what might have caused differences between the replication and the original: *If the models cannot be matched, what causes the differences between them?* The causes for differences in the earlier versions of the replication have been explained above. Even though the majority of experiments are distributionally equivalent to the original, there also existed differences between the original and the final version of the replication. As explained in chapter three, section nine, I believe there are multiple possible causes for these differences. Firstly, the existence of errors in the replication that were not noticed by me when reviewing it. Secondly, the slight possibility of differences between the model as published and the one which was used in the papers by Brughmans and Poblome, or for experiment 26, differences in independent variable values. And thirdly, inherent differences between the ABM toolkits used, which might cause differences in output data, as has been shown in the past.

The fourth question pertains to my critiques of MERCURY: *What consequences, if any, will this replication attempt have on the original study by Brughmans and Poblome* (2016a; 2016b)*?* As discussed in chapter four, through the process of replication, I engaged with the model on a detailed level, which resulted in several methodological criticisms. These included: the inconsistent application of commercial information by traders, the hypothesised network inherently favouring Temin's (2012) hypothesis over Bang's (2008), the fact that the range calculations employed by Brughmans and Poblome are not able to properly visualise the sought after distribution pattern, the limited amount of experiments performed with the network variable that resulted in a tableware distribution close to the archaeological data, and that the unequal distribution of traders among production sites causes great disparities in production quantity. Alternative ways of calculating range distribution were created, which, in addition to new experiments, resulted in data that did not clearly support Temin's (2012) hypothesis over Bang's (2008).

The final research question deals with the relation of my replication study to other one's: *How does this replication of MERCURY compare to other replication studies?* Some issues of replication found by others also arose in this study. Defining what passes as a 'successful' replication has not been clearly defined in the ABM community. As such, when some data sets can be matched statistically while others cannot, there is no consensus on whether this counts as 'successful'. Another common issue is that minor assumptions made in the technical aspects of the model, related to timing or agent selection, can result in significant differences between the replicated model and the original. This problem is often caused by imprecise model descriptions, which was also the case in this study.

By replicating a single model, and comparing my results with other replication studies, I hope to have shown the importance of replication in a broad context. Through replication of agent-based models, researchers cannot only find discrepancies between the description of a model and the way in which it is actually coded, but also errors in the source code, and in my view, more importantly, it allows us to critically engage with the model and critique it on the validation and verification level. Because of these reasons, I believe replicating agent-based models is crucial, both within the field of archaeology as outside of it.

Abstract

This thesis concerns the importance of replication studies in agent-based modelling, specifically in the field of archaeology. As a case study, the MERCURY model by Brughmans and Poblome is replicated.

In the first chapter, a background is given to ABM in general, as well as to replication and its importance and scarcity. Replication allows us to confirm the findings of existing ABM models, or reject them.

The second chapter gives an abstract of Brughmans and Poblome's research. It includes the archaeological background to their research, a precise description of the MERCURY model and a summary of Brughmans and Poblome's conclusions.

In chapter three, the process of replicating MERCURY is explained. Each version of the replication is described in great detail. This final version is statistically compared to the original model. The replication was found to be, for the most part, statistically equivalent to the original. The source of the despondencies between the various versions of the replication and the original model were due to inaccuracies in the description of the model as well as due to my own coding mistakes.

Chapter four includes a brief discussion of existing criticism of the MERCURY model, as well as my own critiques. These critiques are mostly concern the details of the model and the way the authors interpreted their data. Additional experiments are performed to complement the experiments in the original study. I conclude that some of the issues I identify could weaken the original authors' conclusions.

The relation of my replication of MERCURY to other replication studies is discussed in chapter five.

The final chapter my research questions are answered. I also shortly discuss how my experiences with replication could help future researches who want to publish agent-based models or replication studies.

Internet Pages

http://jung.sourceforge.net/doc/api/edu/uci/ics/jung/algorithms/cluster/WeakComponentClusterer.html, accessed on 8 August 2018.

http://repast.sourceforge.net/docs/api/relogo/repast/simphony/relogo/Utility.html#randomExponential(Number), a, accessed on 19 July 2018.

http://repast.sourceforge.net/docs/api/repastjava/repast/simphony/space/graph/ShortestPath.html, b, accessed on 3 July 2018.

https://ccl.northwestern.edu/netlogo/, a, accessed on 4 December 2018.

https://ccl.northwestern.edu/netlogo/5.0/docs/dictionary.html#random-exponential, b, accessed on 19 July 2018.

https://github.com/NetLogo/Network-Extension, accessed on 14 June 2018.

https://libreoffice.org, accessed on 2 December 2018.

https://repast.github.io/, a, accessed on 4 December 2018.

https://repast.github.io/docs.html, b, accessed on 15 May 2018.

https://sourceforge.net/p/repast/mailman/message/36359394/, accessed on 21 August 2018.

https://www.comses.net/codebases/4347/releases/1.1.0/, a, accessed on 16 April 2018.

https://www.comses.net/codebases/d67fd7ce-a6df-4d25-b10c-765b455b80f0/releases/1.0.0/, b, accessed on 5 December 2018.

https://www.jamovi.org/, accessed on 4 December 2018.

Bibliography

Angourakis, A., B. Rondelli, S. Stride, X. Rubio-Campilo, A. Balbo, A. Torrano, V. Martinez, M. Madella and J. Gurt, 2014. Land use patterns in Central Asia. Step 1: The musical chairs model. *Journal of Archaeological Method and Theory* 21(2), 405–425. https://doi.org/10.1007/s10816-013-9197-0

Arikan, B., 2017. Crisis in the Highlands. Agent-based modeling of the Early Bronze Age-I (ca. 4950-4700 cal. BP) Socio-economic transformations at Arslantepe (Eastern Anatolia), in T. Cunningham and J. Driessen (eds), *Crisis to collapse: the archaeology of social breakdown, AEGIS 11*. Louvain: UCL (Université catholique de Louvain) Press, 235–250.

Axtell, R., R. Axelrod, J. Epstein and M. Cohen, 1996. Aligning simulation models: A case study and results. *Computational and Mathematical Organization Theory* 1 (2), 123–141. https://doi.org/10.1007/BF01299065

Axtell, R., J. Epstein, J. Dean, G. Gumerman, A. Swedlund, J. Harburger, S. Chakravarty, R. Hammond, J. Parker and M. Parker, 2002. Population growth and collapse in a multiagent model of the Kayenta Anasazi in Long House Valley. *Proceedings of the National Academy of Sciences of the United States of America* 99 (3), 7275–7279. https://doi.org/10.1073/pnas.092080799

Bajracharya, K. and R. Duboz, 2013. Comparison of three agent-based platforms on the basis of a simple epidemiological model (WIP), in G. Wainer, P. Mosterman, G. Zacharewicz and F. Barros (eds), *Proceedings of the Symposium on Theory of Modeling & Simulation - DEVS Integrative M&S Symposium, San Diego, April 7-10 2013*. San Diego (CA): Society for Computer Simulation International, 1–6.

Bang, P., 2008. *The Roman Bazaar: A Comparitive Study of Trade and Markets in a Tributary Empire*. Cambridge: Cambridge University Press.

Barceló, J., F. Del Castillo Bernal, R. Del Olmo, L. Mameli, F. Quesada, D. Poza and X. Vilà, 2014. Social interaction in hunter-gatherer societies: Simulating the consequences of cooperation and social aggregation. *Social Science Computer Review* 32(3), 417–436. https://doi.org/10.1177/0894439313511943

Barton, C., I. Ullah and S. Bergin, 2010. Land use, water and Mediterranean landscapes: Modelling long-term dynamics of complex socio-ecological systems. *Philosophical Transactions of the Royal Society A* 368(1931), 5275–5297. https://doi.org/10.1098/rsta.2010.0193

Bes, P. and J. Poblome, 2008. (Not) see the wood for the trees? 19,700+ sherds of sigillata and what we can do with them... *Rei Cretariæ Romanæ Fautores Acta* 40, 505–514.

Brantingham, P., 2003. A neutral model of stone raw material procurement. *American Antiquity* 68(1), 487–509. https://doi.org/10.2307/3557105

Briz i Godino, I., J. Santos, J. Galán, J. Caro, M. Álvarez and D. Zurro, 2014. Social cooperation and resource management dynamics among late hunter-fisher-gatherer societies in Tierra del Fuego (South America). *Journal of Archaeological Method and Theory* 21(2), 343–363. https://doi.org/10.1007/s10816-013-9194-3

Brughmans, T. and J. Poblome, 2016a. MERCURY: an agent-based model of tableware trade in the Roman East. *Journal of Artificial Societies and Social Simulation* 19 (1), https://doi.org/10.18564/jasss.2953.

Brughmans, T. and J. Poblome, 2016b. Roman bazaar or market economy? Explaining tableware distributions through computational modelling. *Antiquity* 90 (350), 393–408. https://doi.org/10.15184/aqy.2016.35

Brughmans, T. and J. Poblome, 2017. The case for computational modelling of the Roman economy: a reply to Van Oyen. *Antiquity* 91 (359), 1364–1366. https://doi.org/10.15184/aqy.2017.166

Callegari, S., J. Weissmann, N. Tkachenko, W. Petersen, G. Lake, M. Ponce de León and C. Zollikofer, 2013. An agent-based model of human dispersals at a global scale. *Advances in Complex Systems* 16(4-5), 1350023 (21 pages). https://doi.org/10.1142/S0219525913500239

Cioffi-Revilla, C., W. Honeychurch and J. Rogers, 2015. MASON hierarchies: A long-range agent model of power, conflict, and environment in Inner Asia, in J. Bemmann (ed), *The Complexity of Interaction Along the Eurasian Steppe Zone in the First Millennium CE*. Bonn: Bonn University Press, 89–113.

Clark, J. and S. Crabtree, 2015. Examining social adaptations in a volatile landscape in Northern Mongolia via the agent-based model Ger Grouper. *Land* 4, 157–181. https://doi.org/10.3390/land4010157

Cockburn, D., S. Crabtree, Z. Kobti, T. Kohler and R. Bocinsky, 2013. Simulating social and economic specialization in small-scale agricultural societies. *Journal of Artificial Societies and Social Simulation* 16(4). https://doi.org/10.18564/jasss.2308

Crabtree, S., 2016. Simulating littoral trade: Modeling the trade of wine in the Bronze to Iron Age transition in Southern France. *Land* 5(1), 5 (20 pages). https://doi.org/10.3390/land5010005

Crabtree, S., R. Bocinsky, P. Hooper, S. Ryan and T. Kohler, 2017. How to make a polity (in the central Mesa Verde region). *American Antiquity* 82(1), 71–95. https://doi.org/10.1017/aaq.2016.18

Cuthbert, M., T. Gleeson, S. Reynolds, M. Bennett, A. Newton, C. McCormack and G. Ashley, 2017. Modelling the role of groundwater hydro-refugia in East African hominin evolution and dispersal. *Nature Communications* 8, 15696 (11 pages). https://doi.org/10.1038/ncomms15696

Davies, B. and S. Bickler, 2015. Sailing the simulated seas: A new simulation for evaluating prehistoric seafaring, in A. Traviglia (ed), *Across Space and Time: Papers from the 41st Conference on Computer Applications and Quantitative Meth-*

ods in Archaeology, Perth, 25-28 March 2013. Amsterdam: Amsterdam University Press, 215–223.

Davies, B., S. Holdaway and P. Fanning, 2015. Modelling the palimpsest: An exploratory agent-based model of surface archaeological deposit formation in a fluvial arid Australian landscape. *The Holocene* 26(3), 450–463. https://doi.org/10.1177%2F0959683615609754

Dean, J., G. Gumerman, J. Epstein, R. Axtell, A. Swedlund, M. Parker and S. McCarroll, 2000. Understanding Anasazi culture change through agent based modeling, in T. Kohler and G. Gumerman (eds), *Dynamics in Human and Primate Societies: Agent-Based Modeling of Social and Spatial Processes*. Oxford: Oxford University Press, 179–205.

Donkin, E., P. Dennis, A. Ustalakov, J. Warren and A. Clare, 2017. Replicating complex agent based models, a formidable task. *Environmental Modelling and Software* 92, 142–151. https://doi.org/10.1016/j.envsoft.2017.01.020

Edmonds, B. and D. Hales, 2002. Replication, replication, replication - some hard lessons from model alignment. *Journal of Artificial Societies and Social Simulation* 6 (4), http://jasss.soc.surrey.ac.uk/6/4/11.html.

Epstein, J. and R. Axtell, 1996. *Growing Artificial Societies: Social Science from the Bottom Up*. Brooklyn (NY): Booking Institution Press.

Ewert, U. and M. Sunder, 2018. Modelling maritime trade systems: Agent-based simulation and medieval history. *Historical Social Research* 43(1), 110–143. https://doi.org/10.12759/hsr.43.2018.1.110-143

Festa, P., 2006. Shortest path algorithms, in M. Resende and P. Pardalos (eds), *Handbook of Optimization in Telecommunications*. New York City (NY): Springer, 185–210. https://doi.org/10.1007/978-0-387-30165-5

Goldstein, J., 1999. Emergence as a construct: History and issues. *Emergence* 1, 49–72. https://doi.org/10.1207/s15327000em0101_4

Grimm, V., U. Berger, F. Bastiansen, S. Eliassen, V. Ginot, J. Giske, J. Goss-Custard, T. Grand, S. Heinz, G. Huse, A. Huth, J. Jepsen, C. Jørgensen, W. Mooij, B. Müller, G. Pe'er, C. Piou, S. Railsback, A. Robbins, M. Robbins, E. Rossmanith, N. Rüger, E. Strand, S. Souissi, R. Stillman, R. Vabø, U. Visser and D. DeAngelis, 2006. A standard protocol for describing individual-based and agent-based models. *Ecological Modelling* 198, 115–126. https://doi.org/10.1016/j.ecolmodel.2006.04.023

Grimm, V., U. Berger, D. DeAngelis, J. Polhill, J. Giske and S. Railsback, 2010. The ODD protocol: A review and first update. *Ecological Modelling* 221, 2760–2768. https://doi.org/10.1016/j.ecolmodel.2010.08.019

Gumerman, G., A. Swedlund, J. Dean and J. Epstein, 2003. The evolution of social behavior in the prehistoric American Southwest. *Artificial Life* 9 (4), 435–444. https://doi.org/10.1162/106454603322694861

Heath, B. and R. Hill, 2010. Some insights into the emergence of agent-based modelling. *Journal of Simulation* 4, 163–169. https://doi.org/10.1057/jos.2010.16

Hermellin, E. and F. Michel, 2017. Complex flocking dynamics without global stimulus, in C. Knibbe, G. Beslon, D. Parsons, D. Misevic, J. Rouzaud-Cornabas, N. Brecèche, S. Hassas, O. Simonin and H. Soula (eds), *Proceedings of ECAL 2017: the European Conference on Artificial Life, Lyon, 4-8 September 2017*. Cambridge (MA): MIT (Massachusetts Institute of Technology) Press, 513–520. https://doi.org/10.7551/ecal_a_083

Holland, J., 1995. *Hidden Order: How Adaptation Builds Complexity*. Cambridge: Helix Books.

Huggett, J., 2004. Archaeology and the new technological fetishism. *Archeologia e Calcolatori* 15, 81–92.

Janssen, M., 2009. Understanding Artificial Anasazi. *Journal of Artifical Societies and Social Simulation* 12 (4), http://jasss.soc.surrey.ac.uk/12/4/13.html.

Jin, E., M. Girvan and M. Newman, 2001. Structure of growing social networks.

Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics 64 (4 pt 2). https://doi.org/10.1103/PhysRevE.64.046132

König, D., P. King, G. Laforge, H. D'Arcy, C. Champeau, E. Prag, J. Skeet and J. Gosling, 2015. *Groovy in Action, Second Edition*. Greenwich (CT): Manning Publications.

Kowarik, K., H. Reschreiter and G. Wurzer, 2012. Modelling prehistoric mining. *IFAC Proceedings Volumes* 45 (2), 17–29. https://doi.org/10.3182/20120215-3-AT-3016.00005

Lake, M., 2014a. Explaining the past with ABM: On modelling philosophy, in G. Wurzer, K. Kowarik and H. Reschreiter (eds), *Agent-based Modeling and Simulation in Archaeology*. New York City (NY): Springer, 3–35. https://doi.org/10.1007/978-3-319-00008-4_1

Lake, M., 2014b. Trends in archaeological simulation. *Journal of Archaeological Method and Theory* 21 (2), 258–287. https://doi.org/10.1007/s10816-013-9188-1

Legendi, R. and L. Gulyás, 2012. Replication of the macroABM model: Replication issues in the context of economic agents, in *17th Annual Workshop on Economic Heterogeneous Interacting Agents, Paris, June 21-23 2012*. 15 pages.

Lock, G., 2003. *Using computers in archaeology: towards virtual pasts*. London: Routledge.

Macal, C., 2016. Everything you need to know about agent-based modelling and simulation. *Journal of Simulation* 10, 144–156. https://doi.org/10.1057/jos.2016.7

Macal, C. and M. North, 2009. Agent-based modeling and simulation, in M. Rossetti, R. Hill, B. Johansson, A. Dunkin and R. Ingalls (eds), *Proceedings of the 2009 Winter Simulation Conference, Austin, December 13-16 2009*. Piscataway (NJ): Institute of Electrical and Electronics Engineers Press, 101–112. https://doi.org/10.1109/WSC.2009.5429318

Macy, M. and Y. Sato, 2010. The surprising success of a replication that failed. *Journal of Artifical Societies and Social Simulation* 13 (2), http://jasss.soc.surrey.ac.uk/13/2/9.html. https://doi.org/10.18564/jasss.1611

Mao, G. and N. Zhang, 2017. Fast approximation of average shortest path length of directed BA networks. *Physica A* 466, 243–248. http://dx.doi.org/10.1016/j.physa.2016.09.025

Marwick, B., 2017. Computational reproducibility in archaeological research: Basic principles and a case study of their implementation. *Journal of Archaeological Method and Theory* 24(2), 424–450. https://doi.org/10.1007/s10816-015-9272-9

Miodownik, D., B. Cartrite and R. Bhavnani, 2010. Between replication and docking: ”adaptive agents, political institutions, and civic traditions” revisited. *Journal of Artificial Societies and Social Simulation* 13 (3), https://doi.org/10.18564/jasss.1627.

Ozik, J., N. Collier, J. Murphy and M. North, 2013. The ReLogo agent-based modeling language, in R. Pasupahty, S. Kim, A. Tolk, R. Hill and M. Kuhl (eds), *Proceedings of the 2013 Winter Simulation Conference, Washington DC, December 8-11 2013*. Piscataway (NJ): Institute of Electrical and Electronics Engineers Press, 1560–1568. https://doi.org/10.1109/WSC.2013.6721539

Peña, J., 2007. *Roman Pottery in the Archaeological Record*. Cambridge: Cambridge University Press.

Potting, R., J. Perry and W. Powell, 2005. Insect behavioural ecology and other factors affecting the control efficacy of agro-ecosystem diversification strategies. *Ecological Modelling* 182(2), 199–216. https://doi.org/10.1016/j.ecolmodel.2004.07.017

Romanowska, I., 2015a. Agent-based modelling and archaeological hypothesis testing: The case study of the European Lower Palaeolithic, in A. Traviglia (ed), *Across Space and Time. Papers from the 41st Conference on Computer Applications and Quantitative Methods in Archaeology, Perth, 25-28 March 2013*. Amsterdam: Amsterdam University Press, 203–214.

Romanowska, I., 2015b. So you think you can model? a guide to building and evaluating archaeological simulation models of dispersals. *Human Biology* 87 (3), 169–192. https://doi.org/10.13110/humanbiology.87.3.0169

Rouse, L. and L. Weeks, 2011. Specialization and social inequality in Bronze Age SE Arabia: analyzing the development of production strategies and economic networks using agent-based modeling. *Journal of Archaeological Science* 38 (7), 1583–1590. https://doi.org/10.1016/j.jas.2011.02.023

Rubio-Campillo, X., J. Cela and F. Hernàndez Cardona, 2012. Simulating archaeologists? using agent-based modelling to improve battlefield excavations. *Journal of Archaeological Science* 39(2), 347–356. https://doi.org/10.1016/j.jas.2011.09.020

Santos, J., M. Pereda, D. Zurro, M. Álvarez, J. Caro, J. Galán and I. Briz i Godino, 2015. Effect of resource spatial correlation and hunter-fisher-gatherer mobility on social cooperation in Tierra del Fuego. *PLOS One* 10(4), e0121888 (29 pages). https://doi.org/10.1371/journal.pone.0121888

Scherjon, F., 2012. SteppingIn - modern humans moving into Europe - implementation, in G. Earl, T. Sly, A. Chrysanthi, P. Murrieta-Flores, C. Papadopoulos, I. Romanowska and D. Wheatley (eds), *Proceedings of the 40th Conference on Computer Applications and Quantitative Methods in Archaeology, Southampton, 26–30 March 2012*. Amsterdam: Amsterdam University Press, 105–117.

Shamir, A., 1981. On the generation of cryptographically strong pseudo-random sequences, in S. Even and O. Kariv (eds), *8th International Colloquium on Automata Languages & Programming, Acre, July 13-17 1981*. Berlin Heidelberg: Springer-Verlag, 544–550. https://doi.org/10.1007/3-540-10843-2_43

Temin, P., 2012. *The Roman Market Economy*. Princeton (NJ): Princeton University Press.

Turchin, P., T. Currie, E. Turner and S. Gavrilets, 2013. War, space, and the evolution of Old World complex societies. *Proceedings of the National Academy* *of Sciences of the United States of America* 110(41), 16384–16389. https://doi.org/10.1073/pnas.1308825110

Van Oyen, A., 2017. Agents and commodities: a response to Brughmans and Poblome (2016) on modelling the Roman economy. *Antiquity* 91 (359), 1356–1363. https://doi.org/10.15184/aqy.2017.138

Watts, D. and S. Strogatz, 1998. Collective dynamics of ’small-world’ networks. *Nature* 393, 440–442. http://dx.doi.org/10.1038/30918

Wilensky, U. and W. Rand, 2007. Making models match: Replicating an agent-based model? *Journal of Artificial Societies and Social Simulation* 10 (4), http://jasss.soc.surrey.ac.uk/10/4/2.html.

Will, O., 2009. Resolving a replication that failed: News on the Macy & Sato model. *Journal of Artificial Societies and Social Simulation* 12 (4), http://jasss.soc.surrey.ac.uk/12/4/11.html.

Wobst, H., 1974. Boundary conditions for Paleolithic social systems: A simulation approach. *American Antiquity* 39 (2), 147–178. https://doi.org/10.2307/279579

List of Figures

1 A graph showing the amount of sites each tableware type was found on, based on ICRATES data (Brughmans and Poblome 2016a). . . 18

2 Two example views of the MERCURY world, created using the original NetLogo model (www.comses.net, a). The red dots portray non-production sites and the blue dots portray production sites. The grey lines represent inter-site links. Links between traders on the same site are not visible, since all traders on the same site are occupying the same location. The top image was created using a *proportion-inter-site-links* value of 0,0001 and the bottom one using a value of 0,001. 24

3 Boxplots of the width of distribution, ranked from most to least widely distributed, from three experiments with different *proportion-inter-site-links* values. All other independent variables are the same. This graph was created from data of 100 iterations per experiment (after Brughmans and Poblome 2016b, 403). 29

4 Boxplots of the width of distribution, ranked from most to least widely distributed, from experiments 33 and 34. Both experiments include a disproportional distribution of traders among produciton-sites, i.e. *traders-distribution-site* = '30,1,1,1', but the first experiment was created using the hypothesised network and the second using a random network. All other independent variables are the same. This graph was created from data of 100 iterations per experiment (after Brughmans and Poblome 2016b, 405). 31

5 A bar graph showing the mean clustering coefficient of all 100 iterations per experiment of version one of the replication next to the clustering coefficient of seed 10 of the original model, sorted by experiment number. 37

6 A bar graph showing the mean clustering coefficient of all 100 iterations per experiment of version two of the replication next to the clustering coefficient of seed 10 of the original model, sorted by experiment number. 40

7 A bar graph showing the mean clustering coefficient of all 100 iterations per experiment of version three of the replication next to the clustering coefficient of seed 10 of the original model, sorted by experiment number. 43

8 A bar graph showing the mean clustering coefficient of all 100 iterations per experiment of version four of the replication next to the clustering coefficient of seed 10 of the original model, sorted by experiment number. 47

9 A bar graph showing the mean clustering coefficient of all 100 iterations per experiment of version five of the replication next to the clustering coefficient of seed 10 of the original model, sorted by experiment number. 51

10 A bar graph showing the mean distribution width of all 100 iterations per experiment from version five of the replication next to the same measure from the original model, sorted by experiment number. . . 53

11 A bar graph showing the mean distribution range of all 100 iterations per experiment from version five of the replication next to the same measure from the original model, sorted by experiment number. . . 53

12 A bar graph showing the mean distribution width of all 100 iterations per experiment from version six of the replication next to the same measure from the original model, sorted by experiment number. . . 55

13 A bar graph showing the mean distribution range of all 100 iterations per experiment from version six of the replication next to the same measure from the original model, sorted by experiment number. . . 56

14 A bar graph showing the mean distribution width of all 100 iterations per experiment from version seven of the replication next to the same measure from the original model, sorted by experiment number. . . 58

15 A bar graph showing the mean distribution range of all 100 iterations per experiment from version seven of the replication next to the same measure from the original model, sorted by experiment number. . . 58

List of Tables

1 Independent variables (after Brughmans and Poblome 2016a, supplement 2). 20

2 Dependent variables (after Brughmans and Poblome 2016a, supplement 2). 21

3 A list of all experiments with their independent variables. This table excludes independent variables that are equal across all experiments (after Brughmans and Poblome 2016a, supplement 1). 27

4 This table shows whether the network measures of seed 10 from the original study (Brughmans and Poblome 2016a) are greater than or equal to the minimum and smaller than or equal to the maximum of the same network measures of version 5 of the replication. Note that this does not include the first two steps of network creation, as they always create the same amount of links in both the replication and the original. Link amounts for step three and four were not counted in experiments with randomly created networks in the original, which results in missing data: ’na’. 50

5 The results of Mann-Whitney U tests of tableware average width from the original model versus version 8 of the replication (n=100 for every test). 62

6 The results of Mann-Whitney U tests of tableware range values from the original model versus version 8 of the replication (n=100 for every test). 62

7 The results of Mann-Whitney U tests of tableware average width from random seeds 1-50 versus random seeds 51-100 of the original model by Brughmans and Poblome (2016a) (n=50 for every test). The raw tableware data was shared by Tom Brughmans via email (appendix 1). 63

8 The results of Mann-Whitney U tests of tableware range values from random seeds 1-50 versus random seeds 51-100 of the original model by Brughmans and Poblome (2016a) (n=50 for every test). The raw tableware data was shared by Tom Brughmans via email (appendix 1). 63

9 A comparison of range values of experiments 1 to 12 as calculated by Brughmans and Poblome (2016a), by percentage difference between the most and least widely distributed product, the most minus the second most widely distributed product and the percentage difference between the most and second most widely distributed product. 72

10 A comparison of range values of experiments 24 to 28 and 33 as calculated by Brughmans and Poblome (2016a), by percentage difference between the most and least widely distributed product, the most minus the second most widely distributed product and the percentage difference between the most and second most widely distributed product. 74

11 Average product width and ranges of the new experiments 36 to 40 and old experiment 33 for comparison. In these experiments, *proportion-inter-site-links* is varied. All other independent variables are kept constant: *equal-traders-production-site* = ’false’, *traders-distribution* = ’exponential’, *network-structure* = ’hypothesis’, *local-knowledge* = 0,5, *traders-production-site* = ’30,1,1,1’, *max-demand* = 10. 74

List of Appendices

Appendix 1: Raw tableware data of the original MERCURY model, courtesy of Tom Brughmans.

Appendix 2: Repast source code of version 1 of the replication.

Appendix 3: Output data of version 1 of the replication.

Appendix 4: Repast source code of version 2 of the replication.

Appendix 5: Output data of version 2 of the replication.

Appendix 6: Repast source code of version 3 of the replication.

Appendix 7: Output data of version 3 of the replication.

Appendix 8: Repast source code of version 4 of the replication.

Appendix 9: Output data of version 4 of the replication.

Appendix 10: Repast source code of version 5 of the replication.

Appendix 11: Output data of version 5 of the replication.

Appendix 12: Repast source code of version 6 of the replication.

Appendix 13: Output data of version 6 of the replication.

Appendix 14: Repast source code of version 7 of the replication.

Appendix 15: Output data of version 7 of the replication.

Appendix 16: Repast source code of version 8 of the replication.

Appendix 17: Output data of version 8 of the replication.

Appendix 18: Output data of the new experiments, using version 8 of the replication.

Appendix 19: Excerpt of my email correspondence with Tom Brughmans.