Modelling bivariate exposure distributions for Out of Home advertising

Modelleren van bivariate blootstellingsverdelingen voor Out of Home reclame

by

Frederiek Backers

A thesis submitted to the Delft University of Technology to obtain the degree of Bachelor of Science at the Delft University of Technology, to be defended publicly on Friday July 8th, 2022 at 01:00 PM.

Student number:	4704452	
Thesis committee:	Dr. ir. G.F. Nane	TU Delft (supervisor)
	Dr. A. Bishnoi	TU Delft
Internship supervisors:	T. Butterbrod	Greenhouse Group
	F. van den Berg	Greenhouse Group

TU Delft

Greenhouse

Abstract

Out of Home advertising is traditional outdoor advertising. Over the last decade the digital Out of Home advertising possibilities have grown substantially. These possibilities include hour-based advertisement schedules, rapidly implemented changes to advertisements and obtaining location data from consumers being exposed to advertisements. Investigating these possibilities further for improving customer engagement and optimising target group reach is of great importance in marketing and marketing science. This thesis will focus on exposure data obtained from vehicle locations in central London. Using this data we want to model the frequency with which individuals see advertisements and research the overlap in exposures of individuals to multiple media vehicles where advertisements are displayed. We focus on modelling exposures of individuals to two vehicles, comparing different types of bivariate distributions. In this thesis we present the comparison of two models: an adapted version of the Sarmanov bivariate distribution and copulas. We refer to these models as the Danaher and Copula model respectively.

The fitting and simulations of the models are based on bivariate data of two specified vehicles, to be able to comment on the overlap between these vehicles. To investigate the performance of the models all combinations of vehicles are modelled after which the results are combined. Both models simulate small frequencies of exposures of individuals adequate, but for more extreme values both models fail to represent the observed data. Especially the simulation of overlap is done poorly by the models. Several approaches have been tried to improve the models' performance, without much success. The analysis shows that the data at hand requires more sophisticated methods to handle joint exposure and more research needs to be done.

In addition the variable distance is added to the research problem, which intuitively has great influence on the overlap between media vehicles. This is done by using 3-dimensional copulas. For this research we modify the data: instead of first fitting and simulating a copula model on the data of two vehicles and then combining the results of all vehicle combinations, we first combine the data from all the combinations of vehicles and repeat the fitting procedure thereafter. Several options for the modelling of these copulas are presented. Similar analysis shows, again, that the nature of the data requires more complex models to handle the overlap while also accounting for the vehicle distance. Finally, using the modified data we look back at the 2-dimensional copula model and perceive an excellent resemblance of the data for the Student-t copula.

Preface

This thesis has been written in order to obtain the degree of Bachelor of Science. The writing and research has been conducted under supervision of Dr. ir. G. F. Nane from the department of Applied Probability of the faculty EEMCS at Delft University of Technology.

I would like to thank Tina for the support in the process of writing this thesis. Thank you for your patience, trust and half-full glasses motivational speeches which helped me in difficult times. Your ideas and supervising approach ensured I made it to the other side.

Furthermore, I want to thank some people from Greenhouse Group that enabled me to do this thesis project: Ruben Mak and Bart Frenk. Thank you for the opportunity to conduct this research as an intern at Greenhouse, it was challenging at times but very educational and fun. Especially thanks to Fabian van den Berg and Tim Butterbrod, who sticked with me till the end and showed me there is more to research than just sitting behind a dusty computer. Thank you for all your time and supervision, I am very grateful for your encouragement.

Thank you Bastiaan, Vera, Sterre, Kilian, Max, Koen and Sophie for your friendship and support these last couple of months.

Lastly I want to thank Anurag Bishnoi for taking place in my thesis examination committee.

Frederiek Backers
Delft, July 2022

Contents

1

Introduction

One of the most used forms of advertising is Out Of Home (OOH) advertising. These are offline campaigns on for example billboards and bus shelters that give brands the possibility to communicate with a large number of consumers. Posters used to be the custom OOH advertisement, where a company could buy a spot for a specific amount of days. Since the beginning of the digital era, these poster spots are gradually being replaced by digital media vehicles, which allows companies to buy spots per hour and make more decisions based on data obtained by these vehicles. Companies have the options to buy advertisement spots on media vehicles that are placed on fixed locations throughout cities and countries. A good advertisement reaches as many target group individuals as possible, while minimising costs for advertisement spots. The **reach** of an advertisement is the proportion of the target audience exposed to at least one ad. To minimise the costs, we want to look at the frequencies with which we reach the audience. The **frequency** is the number of exposures an individual has to an ad. The most efficient/cost-effective media strategy would be to reach as many unique individuals as possible whilst keeping the number of individuals seeing multiple vehicles, i.e. overlap, to a minimum. The underlying problem is therefore: modelling the frequencies of individuals to media vehicles and modelling the overlap between vehicles.

To solve this problem we have data available from media vehicles scattered around the United Kingdom. This data consists of exposures of individuals to media vehicles[1]. A thorough explanation of the data is given in chapter 2. Using this data, we can try to model the exposures an individual has to one vehicle, two vehicles, three vehicles, etc. This means modelling a multivariate exposure distribution. Using this distribution, we can model the frequency that individuals have to an ad and estimate the overlap of reached individuals between vehicles. In this thesis we will focus on modelling a bivariate distribution, trying to model the overlap between different combinations of two vehicles.

We will try two different models for the modelling of the bivariate exposure distribution: the Danaher model and Copula model. These two models both have significant presence within marketing science and previous similar problems, which is why we will center this thesis on these two models. In chapter 3 we will discuss the Danaher model, a model that is based on the Sarmanov bivariate distributions, which consists of univariate marginals and a correction factor that describes correlation between vehicles. The model development consists of two parts: the one-vehicle model, where we fit a distribution to the univariate data of vehicles, and the two-vehicle model, where we build a special version of the Sarmanov bivariate distribution using the one-vehicle model. We will show the estimation of parameters, simulate from our models, compare this to the data and show the results. In chapter 4 we will discuss some copula models, which are models that can describe dependencies between random variables. We will give theoretical background on the most used copulas and their parameters. Afterwards we will fit different copulas and do simulations to see which copula model represents our data best. In chapter 5 we will compare the Danaher and Copula model to see which model is most appropriate for modelling exposures for out of home advertisements.

[1] An 'exposure' in our data means the person moved past the vehicle. We do not know for certain if the person actually saw the vehicle, but we will assume this.

In addition to modelling exposures and overlap between vehicles, we also investigate distance between vehicles. This third variable is added because it intuitively seemed to correlate with the previous two. Two vehicles placed closely together would expectedly show higher overlap, especially if these vehicles are placed in crowded places such as malls or train stations. Therefore, we will explore this idea in chapter 6.

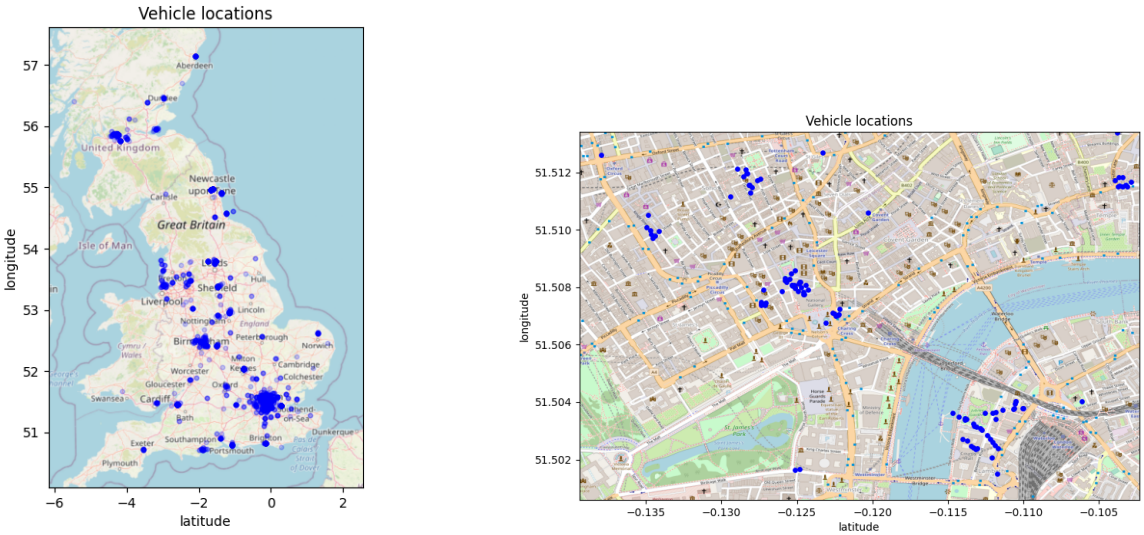
Lastly, we will conclude our findings in chapter 7 and discuss these conclusions and possible recommendations for future research in chapter 8.

2

Data

The data available for fitting and testing our models is obtained from media vehicles stationed in the UK. These media vehicles are OOH sites and store locations, scattered throughout the UK, but clustered around cities and their centers. Throughout the course of 28 days, these media vehicles tracked over 200.000 individuals using GPS. Each data point represents one exposure of an individual to a media vehicle. Therefore, for every vehicle we can identify which individuals were reached and how many times (i.e. frequency) these individuals were reached.

We also have information about the location of each vehicle. Every vehicle has an identity number that can be linked to a postal code. Using these postal codes we can calculate the distance between vehicles. In figure 2.1 all media vehicles in our dataset are marked by a blue dot. It can be seen that the vehicles are clustered around and within cities.



(a) Vehicle locations in UK.

(b) Vehicle locations in London.

Figure 2.1: Locations of media vehicles in available dataset.

Let us focus on individual vehicles and the individuals exposed to these vehicles. In figure 2.2 two examples of vehicles and their frequencies are displayed.

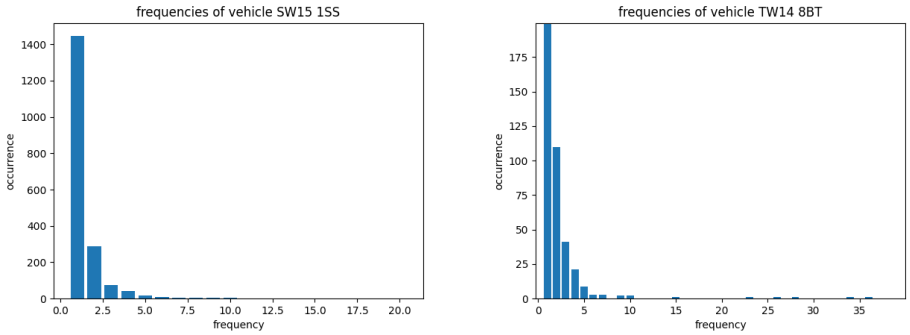


Figure 2.2: Two examples of vehicles and their occurring frequencies.

In these frequency plots we can see how many times individuals have seen the two vehicles *SW15 1SS* and *TW14 8BT*. For the left vehicle around 1400 people saw the vehicle once, around 300 people saw the vehicle twice, about 75 people thrice, etc. On the right, we zoomed in to be able to see the higher frequencies: there are small peaks around frequencies 10, 25 and 35. We can see that there are some individuals that see the vehicle this many times. So the occurrence of frequency 1 is a high peak from where we can see a steep curve going to zero, with some small peaks at high value frequencies. This is expected since the higher the frequency, the more times a person is exposed to this vehicle, which is more unlikely the higher the frequency.

When modelling the overlap between vehicles, it is not sufficient to only model individuals that were reached by vehicles. It is also necessary to model individuals that were not reached by vehicles, i.e., the non-reach. However, obtaining this non-reach data by comparing one vehicle to all other vehicles leads to zero-inflation, since there are generally more vehicles out of overlap reach than within. Zero-inflation is problematic for the models we are attempting to use. In figure 2.3 we compare the data of two combinations of vehicles: two vehicles that have a large overlap (blue), and two vehicles without any overlap (green). For the data we take into account the individuals that have seen at least one of the two vehicles, therefore already removing the zeros created by individuals that have not seen any of the two vehicles.

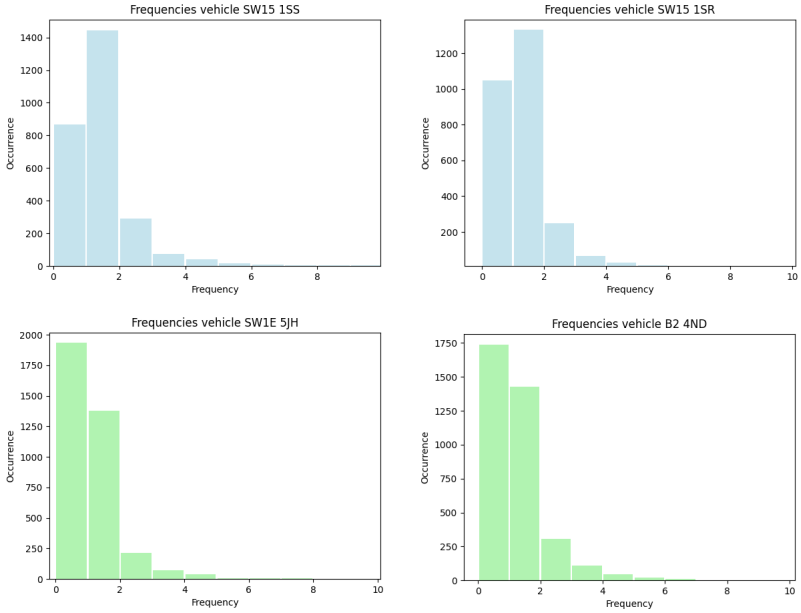


Figure 2.3: Examples of vehicle data of vehicles with large overlap (blue) and no overlap (green).

Figure (2.3) shows that if we compare the data of two vehicles with large overlap (blue) to the data of two vehicles with no overlap (green), the number of zeros increases significantly.

The main goal of this thesis is to model the potential overlap in exposures in the above described data. Overlap between the exposures of two vehicles implies that these are seen by the same individuals, decreasing their effectiveness. As mentioned above, the raw data contains a large amount of zeros. The next section will describe how the data was selected to decrease the number of zeros.

2.1. Selecting data

The previous section already touched upon the problem of zero-inflation in the current dataset. An acceptable solution to this problem would be to limit the dataset based on a key assumption. In addition, limiting the dataset would also improve the efficiency and computation time of the model. The data was therefor selected based on geographic location. We assumed that vehicles from a city such as Glasgow, Scotland would not need to be compared to vehicles from another city such as Manchester, England, as the probability of individuals seeing vehicles at both locations is assumed to be negligible. For the current dataset we selected a region with a reasonably high number of individuals that saw at least one vehicle. The remainder of this thesis will focus on a small region in the center of London. This region consists of 23 vehicles, which are seen by 3612 unique individuals. The locations of the vehicles are displayed in figure 2.4.

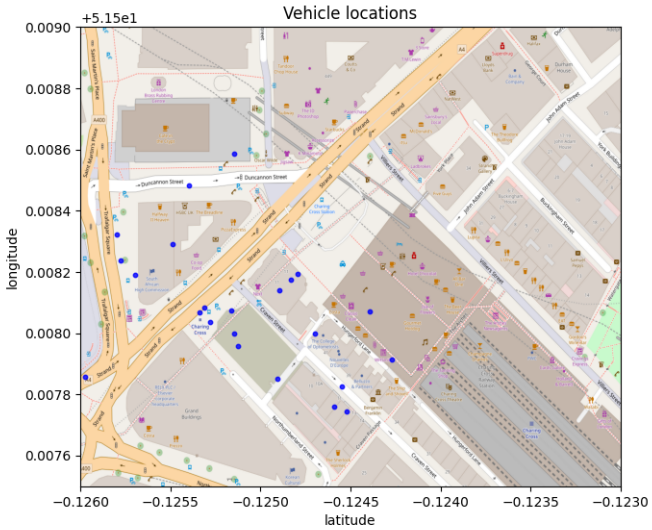


Figure 2.4: Selected vehicle locations in the center of London.

The models investigated in this thesis will be fitted onto the selected data. The simulated data is then compared to the observed data to determine which model shows the best fit. Fitting the models is done on bivariate data: the individual exposures to a combination of two vehicles. An example of this bivariate data can be seen in table 2.1. In this example the the individual exposures of vehicle 1 and 2 create a two-column matrix, where the columns represent the vehicles and the rows represent the individuals of our dataset. Every value in row i and column j represents how many times individual i saw vehicle j. An example of this can be seen in the following table:

	vehicle 1	vehicle 2
person 1	4	2
person 2	1	0
person 3	10	5
person 4	0	0
person 5	0	1
person 6	0	0

Table 2.1: Constructed example data for two vehicles and six individuals.

In the example above, person 1 has seen both vehicles; vehicles 1 four times and vehicle 2 twice. Person 2 had only seen vehicle 1 once, whereas person 4 has neither seen vehicle 1 or 2. Due to the previously mentioned selection criteria we know that person 4 has seen at least one vehicle once, but in this example that is neither vehicle 1 or 2. We had to transform our original data to this format to be able to do calculations in the forthcoming of this thesis. The two-vehicle matrix can be plotted in a 2-dimensional scatterplot with each axis representing one of the two vehicle exposures. Repeating the procedure explained above for every combination of the 23 vehicles in the selected region and combining the results, results in the following scatterplot:

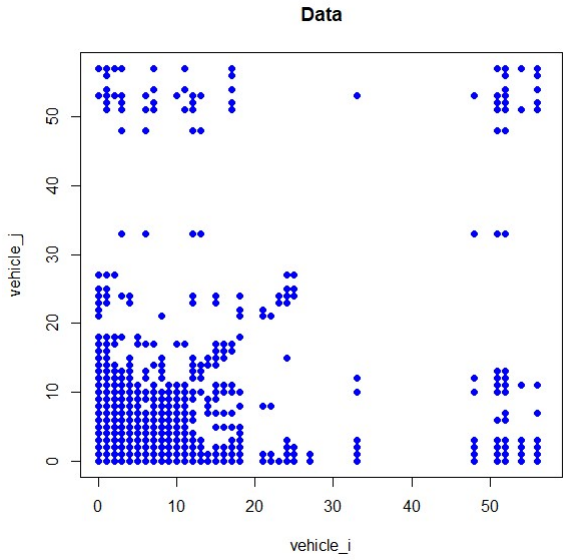


Figure 2.5: Datapoints of our selected data consisting of all possible two-vehicle combinations and their occurring frequencies.

This figure represents all the datapoints that will be used in fitting and comparing the models throughout this thesis. We can see that the datapoints are clustered around certain areas of the plot. We can translate these clusters to the overlap described in chapter 1. The lower left corner represent the combinations of frequencies (i, j) where both i and j are small values. For example a person seeing arbitrary vehicle 1 two times and arbitrary vehicle 2 three times. If neither i nor j are zero these points represent small overlap. If either i or j or both have frequency zero, this means they do not overlap, which we count as non-overlap. The points on the x and y axis represent the non-overlap, also the clusters in the upper left and lower right corner represent non-overlap where one of the two vehicles is seen with a high frequency. Then we also see a small cluster in the upper right corner, which means that both vehicles have been seen a lot of times by the same person. These values can be seen as 'extreme' overlap and are very much of interest for media planning.

2.2. Adding distance

In chapter 6 we will add the variable distance to find a relation between distance and exposures/overlap. We think this variable is of great influence on the overlap between vehicles. This is because we would expect there to be a larger overlap between vehicles if these vehicles are placed closer together. To support this

idea the Kendall tau correlation (which will be explained in more detail in the model chapters), is plotted against the distance in meters between vehicles in figure 2.6. This Kendall tau correlation is a correlation that describes the dependency between random variables. In our case this would mean that if two vehicles are highly positively correlated, the chance of seeing both vehicles becomes higher if you see either one of the two. We can see that the higher the distance, the more the correlation coefficient goes to zero. This is a very interesting result which we will explore more in chapter 6.

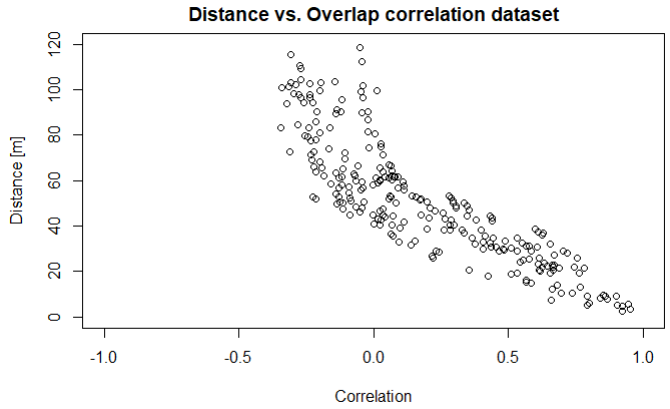


Figure 2.6: The Kendall tau correlation between two vehicles with respect to the distance between the two vehicles.

3

Model 1 - Danaher

3.1. Introduction

The first model we will use for modelling frequencies is the Danaher model. This model by Danaher (2007) [1] is used to model page views across an arbitrary number of websites with an application to reach and frequency prediction. Danahers model outperforms a lot of existing traditional/online media models such as the Tobit model by Li et al (2002) [2] , Leckenby and Hong's(1998) [3] , Huang and Lin's(2006) [4] and Wood's(1998) [5] . This is why we choose this model for our problem.

We start off by modelling the exposures an individual has to a media vehicle by formulating an exposure distribution, the one-vehicle model, and explain how to estimate the parameters used in this model. We then extend this univariate distribution to a bivariate distribution using the Sarmanov bivariate distribution, which is described in section 3.4 and give the method used for estimating the parameters. We will discuss the correlation coefficient between vehicles for this model and show the results that we found.

3.2. Model development

Let X_i be the number of exposures a person has to media vehicle i:

$$X_i = 0, 1, 2, \ldots \; ; \; i = 1, \ldots, m$$

where m is the number of vehicles in the selection of our data. For our dataset, there are 23 vehicles. Over a period of time every X_i can range from 0 to infinity with discrete values since there is no limit for the number of times an individual can pass by a vehicle. For modelling the overlap between vehicles we will model the joint bivariate distribution of (X_i, X_j), for every $i \neq j$ since we do not compare vehicles to themselves and $i < j$ to ignore symmetry, i.e., (X_1, X_2) being the same as (X_2, X_1). In Danahers model we first model the univariate exposure distribution, which we use in the bivariate case.

3.3. One-vehicle model

Danaher argues that modelling the number of exposures a person has to a single website i in a fixed time period is analogous to a problem in marketing science where we need to model the number of purchases a person makes in a product category. This is accurately modelled by a Poisson distribution with rate parameter λ coming from a gamma distribution [15][16]. We are going to use this model for our one-vehicle problem. This compound $X_i \mid \lambda_i \sim Poisson(\lambda_i)$ gives us the Negative Binomial Distribution (NBD) with mass function:

$$P(X_i = x_i \mid r_i, p_i) = \binom{x_i + r_i - 1}{x_i} p_i^{r_i} (1 - p_i)^{x_i} \quad x_i = 0, 1, 2, \ldots \tag{3.1}$$

where r_i and p_i are the parameters for the number of failures before the r'th success and the probability of success in each trial respectively. In our problem we can interpret these parameters as the number of failures before we reach the frequency of interest and the probability of success. We parameterize this mass function

by $\alpha = \frac{p}{1-p}$ (or $\alpha = \frac{r}{mean}$, which is used in R) to obtain the following mass function which we will use in the univariate and bivariate models:

$$P(X_i = x_i \mid r_i, \alpha_i) = \binom{x_i + r_i - 1}{x_i} \left(\frac{\alpha_i}{\alpha_i + 1}\right)^{r_i} \left(\frac{1}{\alpha_i + 1}\right)^{x_i} \quad x_i = 0, 1, 2, ... \tag{3.2}$$

3.3.1. Parameter estimation

For fitting the NBD to our data, we use the *fitdist()* function from the *fitdistrplus* [6] package in R. This function estimates the parameters r and p using Maximum Likelihood Estimation (MLE). We calculate α using the parametrization $\alpha = \frac{p}{1-p}$ (or $\alpha = \frac{r}{mean}$). The MLE fits parameters to the data set, under the assumption that the samples are sampled from the same distribution. In general the distribution parameter θ is estimated by maximizing the likelihood function, which is defined as:

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

with x_i the n observations of random variable X and $f(\cdot|\theta)$ the density/mass function of the parametric distribution. In our case, $f(\cdot|\theta)$ is the mass function of the NBD from equation (3.1). As an example, the maximum likelihood estimation of parameter p for the NBD can be found in appendix A.1.1.

3.4. Two-vehicle model

An extension from our one-vehicle model to a two-vehicle model is not easily derived. In the one-vehicle model the NBD is derived from the marketing problem of modelling purchases within a product category. Extending this to two categories is relatively difficult since we have to take into account dependencies between categories. If we assumed that the different categories, or two media vehicles, are independent from each other, we would simply multiply the two marginal probability mass functions of the NBD and this would give us the bivariate mass function that we are looking for. But we know that there is a certain correlation between media vehicles, i.e. the same person being exposed to multiple vehicles, so we cannot assume independence between the vehicles. In our two-vehicle model we need to account for this "overlap" correlation. For this purpose, Danaher uses the Sarmanov bivariate distribution (3.3). This distribution is developed by Sarmanov in 1966 and first applied by Lee (1996) [2] in statistics literature. The general form of the Sarmanov bivariate distribution for (X_1, X_2) is:

$$f(X_1, X_2) = f_1(X_1) f_2(X_2)[1 + \omega \phi_1(x_1)\phi_2(x_2)] \tag{3.3}$$

where $f_i(X_i = x_i)$ is the marginal distribution for X_i and $\phi_i(x_i)$ are called "mixing functions", which have the requirement that $\int \phi_i(t) f_i(t) dt = 0$. We can see this bivariate distribution as the product of two marginal distributions, allowing for correlation between the distributions with a "correction factor", which is given by the second part of the equation. The two-vehicle model is described by our one-vehicle NBD's together with a correction factor, expressed by the parameter ω and some mixing functions. The estimation of ω is given in section 3.4.1, the mixing functions can be chosen arbitrarily, but there are some examples for certain distributions. Lee (1996) [2] recommends the mixing functions for the NBD to be:

$$\phi_i(x_i) = e^{-x_i} - \left(\frac{\alpha_i}{1 + \alpha_i - e^{-1}}\right)^{r_i} \tag{3.4}$$

where α_i and r_i are the estimated parameters from our one-vehicle model. We will use these mixing functions in our bivariate model. Now substituting the mixing functions of equation (3.4) in the general form of the Sarmanov distribution given in equation (3.3), we obtain the bivariate distribution for our two-vehicle model of (X_1, X_2):

$$f(X_1, X_2) = f_1(X_1) f_2(X_2) \left[1 + \omega \left(e^{-x_1} - \left(\frac{\alpha_1}{1 + \alpha_1 - e^{-1}}\right)^{r_1}\right) \cdot \left(e^{-x_2} - \left(\frac{\alpha_2}{1 + \alpha_2 - e^{-1}}\right)^{r_2}\right)\right] \tag{3.5}$$

where $f_i(X_i)$ are the NBD distributions given by equation (3.2) with parameters α_i and r_i. We now developed a two-vehicle model consisting of the product of univariate NBD's with a correction factor to account for the correlation in exposures between media vehicles.

3.4.1. Parameter estimation

The parameters in the two-vehicle model are α_i, r_i and ω. The estimation method of α_i and r_i are described in section 3.3.1. For the estimation of ω Danaher suggest three different methods: maximum likelihood, method of moments and method of means and zeros. Danaher tested all three methods and it turned out that the method of means and zeros (explained below) is empirically superior to the others. The reason for this difference is in the estimation of reach: the proportion of audience exposed to at least one ad. The method of means and zeros results in a very accurate estimate of reach, which is the most important measure in media planning. This is because the reach is one minus the non-reach, therefore the method of means and zeros ensures an exact match between the model estimate and observed value of the bivariate non-reach. This is why we will use the method of means and zeros to estimate parameter ω and leave the other types of estimators for further research.

The method of means and zeros for the bivariate model (3.5) works as follows: the non-reach for two arbitrary vehicles X_1 and X_2 is:

$$f(X_1 = 0, X_2 = 0) = f_1(X_1 = 0) f_2(X_2 = 0)[1 + \omega\phi_1(0)\phi_2(0)] \tag{3.6}$$

where $f_i(X_i = 0) = f_i(0|r_i, \alpha_i)$ and $\phi_i(0) = 1 - \left(\frac{\alpha_i}{1+\alpha_i - e^{-1}}\right)^{r_i}$, $i = 1,2$. Now obtain the observed non-reach of two vehicles using the estimated $\hat{r}_i$ and $\hat{\alpha}_i$ from section 3.3.1 and then equation (3.6) to obtain the estimate of ω:

$$\hat{\omega} = \left[\frac{\hat{f}(0,0)}{\hat{f}_1(0)\hat{f}_2(0)} - 1\right] \frac{1}{\hat{\phi}_1(0)\hat{\phi}_2(0)} \tag{3.7}$$

This method is used for all pairwise combinations of vehicles, which are $\frac{m(m-1)}{2}$ pairwise combinations of vehicles.

3.5. Correlation

For the generalized Sarmanov distribution (3.3) Lee (1996) [2] gives a general expression for the correlation between two random variables. In our bivariate distribution (3.5), the correlation between two arbitrary vehicles X_1 and X_2 is given by:

$$corr(X_1, X_2) = \omega(1 - e^{-1})^2 \frac{\sqrt{r_1 r_2 (1+\alpha_1)(1+\alpha_2)}}{\alpha_1 \alpha_2} \cdot \left(\frac{\alpha_1}{1+\alpha_1 - e^{-1}}\right)^{r_1+1} \left(\frac{\alpha_2}{1+\alpha_2 - e^{-1}}\right)^{r_2+1} \tag{3.8}$$

From this equation we can see that X_1 and X_2 are uncorrelated if and only if $\omega = 0$, since the parameters from the univariate NBD are unequal to zero for all vehicles. Therefore ω largely determines the correlation between the exposures of two media vehicles.

3.6. Results

We use the selected regional data from London city center as described in section 2.1 to fit and test our model, this data consists of 23 vehicles which gives us $\frac{23(23-1)}{2} = 253$ pairwise combinations of vehicles. The implementation of the model has been done in R and the functions can be found in Appendix A.1.2. Due to the high running time ($\sim$ 30 minutes per combination) of the calculation for the bivariate distribution we do not have enough time to run all different combinations of vehicles. Therefore we randomly picked 44 combinations of vehicles to fit and test the two-vehicle model. The results for the one- and two-vehicle models are presented in sections 3.6.1 and 3.6.2 respectively.

3.6.1. One-vehicle model

In the one-vehicle model we fit the NBD from equation (3.2) to our data using the *fitdist()* function from the *fitdistrplus* package in R [6] which is described in section 3.3.1. In the figure below you can find four examples of randomly chosen vehicles and their NBD fit.

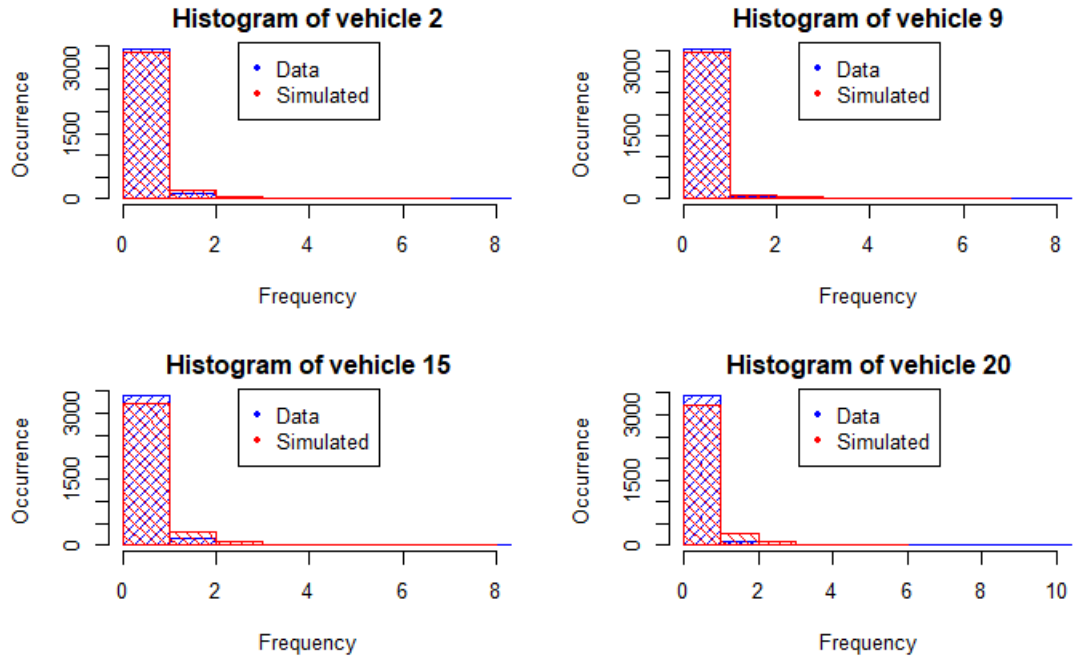


Figure 3.1: Examples for vehicles 2, 9, 15 and 20 of their data compared to the NBD fit of the one-vehicle model.

We can see that the NBD seems to be fitting our data adequately. To test if the NBD fits our data we apply a Goodness-of-Fit test. We choose between two tests appropriate for discrete data: the Chi-Squared test and the exact test of goodness-of-fit. The rule of thumb for the use of Chi-Squared test is that all expected values must be greater than five, which is not the case for our problem since the expected values for high frequencies are very small ($\sim 10^{-20}$). Therefore we use the exact test of goodness-of-fit (*binom.test()* function in R), which is a statistical hypothesis test used to determine if the proportions of the sample space are equal to the expected proportion. We do a two-tailed test, meaning we are stating that the observed proportions are equal to the expected proportions. Our statistical hypothesis with significance level $\alpha = 0.05$ is as follows:

- *Null hypothesis* H_0: There is no significant difference between the data and fitted NBD.
- *Alternative hypothesis* H_1: There is a significant difference between the data and fitted NBD.

For every vehicle we need to individually calculate the p-value for each occurring frequency. We then take the average of all these individual p-values to get an average p-value for the fit of the NBD to our data. Due to lack of time we do this for four randomly picked vehicles out of the 23 vehicles. The resulting p-values can be found in table 3.1.

vehicle number	average p-value
2	0.7703
9	0.7028
15	0.4950
20	0.3659

Table 3.1: Average p-value for the exact test of goodness-of-fit for randomly picked vehicles under the statistical hypothesis stated above.

We can see that every p-value is greater than the significance level of α. Therefore we do not reject the null hypothesis and conclude that the NBD is a reasonable fit to our data.

3.6.2. Two-vehicle model

For the two-vehicle model we have 253 pairwise combinations of vehicles to fit our bivariate distribution from equation (3.5). As stated in the beginning of this section, due to the high running time of the calculation of this distribution, we do not have enough time to run all different combinations. Therefore we randomly

picked 44 combinations of vehicles to fit and test our model. The functions made for the implementation of this model can be found in Appendix A.1.2. The procedure for the two-vehicle model for two arbitrary vehicles 1 and 2 is the following:

1. Fit the univariate NBD to both vehicles using the *fitdist()* function in R.
2. Extract the estimate for parameters $size = r$ and $mu = mean$ from the fits.
3. Using the parametrization $\alpha = \frac{r}{mu}$ obtain the estimate for parameter α.
4. Estimate parameter ω.
5. Calculate the joint distribution using equation (3.5).
6. Make a joint probability table where the rows and columns represent the occurring frequencies of vehicle 1 and 2 respectively.
7. Sample from our model using the probabilities from the probability table.

Using this procedure on the 44 randomly picked pairwise combinations resulted in the following figures for the simulated data plotted with the observed data (figure (3.2)) and for comparison the real data (figure (3.3)).

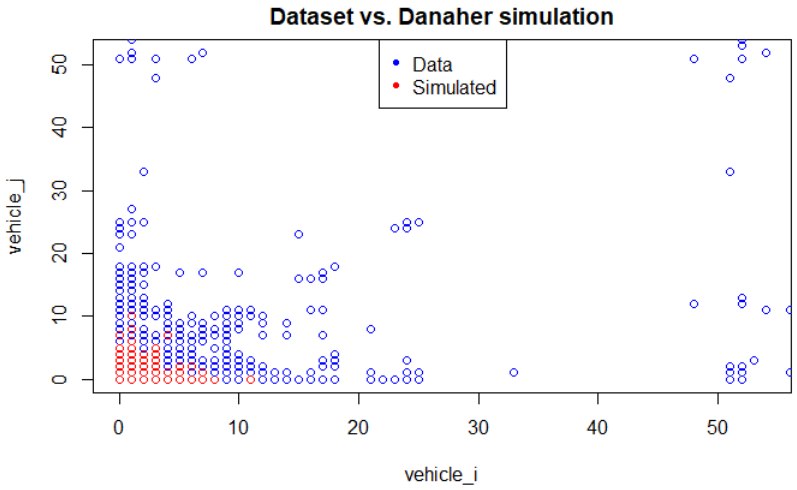


Figure 3.2: Our data and the simulated data from the Danaher two-vehicle model.

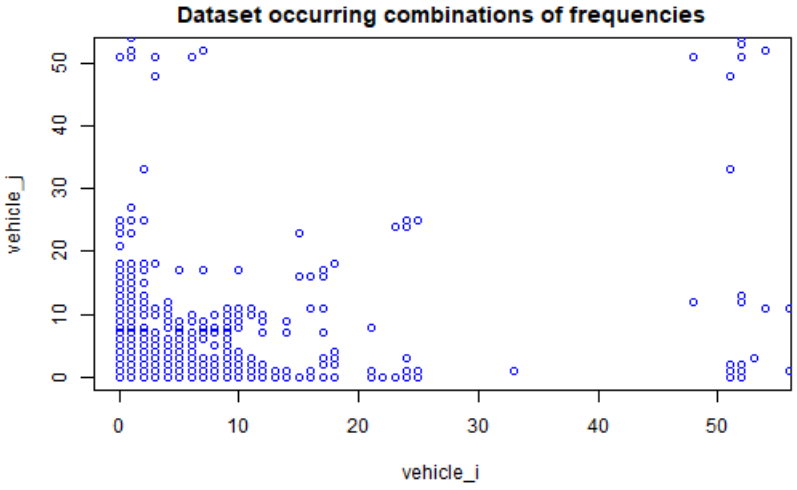


Figure 3.3: Dataset occurring combinations of frequencies for better comparison.

On the x-axis and y-axis we can see the frequencies of the vehicles i and j. Every data point represents an individual and the number of times this individual was exposed to vehicle i and the number of times this individual was exposed to vehicle j. The blue points are the real data and the red points are the simulated data from the two-vehicle model. The red points are plotted over the blue points. We can see that the Danaher model fits the combinations for small values of frequencies, but for larger frequencies the Danaher model fails to simulate any data points. The larger values of frequencies represent the extreme overlap values, where one individual sees one or both of the vehicles a large number of times.

There are no established statistical methods to test the fitness of our bivariate distribution. We chose to adjust the method used in Danaher [1] by looking at the capturing of overlap of the model. We want to see how well the two-vehicle model captures the non-overlap, overlap and extreme overlap (frequencies both vehicles > 4). To do this, for every pairwise combination of vehicles (in this case the 44 randomly picked ones) we look at the occurring combinations of frequencies of every individual and count the non-overlap, overlap and extreme overlap. The non-overlap is the number of individuals that see either vehicle 1, 2 or both vehicles 0 times ((0,0), (i,0), (0,j) : i,j > 0). The overlap is 1 - non-overlap, but also the number of individuals that see both vehicles at least once ((i,j) : i,j > 0). The extreme overlap is the number of individuals that see both vehicles more than four times ((i,j) : i,j > 4). We count the non-overlap, overlap and extreme overlap for all 44 combinations, for the real data and the simulated data from the Danaher model. Next we sum these results to obtain the total points for the data and Danaher model in the three overlap categories. Lastly, we calculate the percentage in which the Danaher model captures the non-overlap, overlap and extreme overlap with respect to the data, to be able to test the fit of the Danaher model. In the table below you can see three vehicle combination examples and their count for the three overlap categories. You can also see the summed values for the non-overlap, overlap and extreme overlap. Lastly, the percentage of the captures is presented.

vehicle combinations	data	Danaher model	type overlap
(1,2)	3277	3299	non-overlap
(2,3)	3467	3480	
(1,3)	3114	3122	
(1,2)	335	313	overlap
(2,3)	145	132	
(1,3)	498	490	
(1,2)	11	1	extreme overlap
(2,3)	0	0	
(1,3)	8	0	
.	.	.	.
.	.	.	.
.	.	.	.
(0,0), (i,0), (0,j) : i,j > 0	142731	154603	non-overlap
(i,j) : i,j > 0	16197	4325	overlap
(i,j) : i,j > 4	334	5	extreme overlap
captured percentage w.r.t. data	100%	108%	non-overlap
	100%	27%	overlap
	100%	1.49%	extreme overlap

Table 3.2: An overview of the captured non-overlap, overlap and extreme overlap for the Danaher model.

The first three blocks of rows are three examples for the number of points that represent non-overlap, overlap and extreme overlap. For example: the combination of vehicle 1 and 2 (represented as (1,2)) the number of points that have no overlap in the data is 3277 and in the simulations from the Danaher model this number is 3299. The number of points that have overlap are 335 for the data and 313 for the Danaher model, for the extreme overlap these numbers are 11 and 0. When we sum all these values for the non-overlap, overlap and extreme overlap over all the different combinations of vehicles (in this case 44), we get the total number of points in the three overlap categories for the data and Danaher model. These numbers can be found in the fifth block of rows. In the last block we presented the captured percentage with respect to the data, which is trivially 100% for the data, but for the Danaher model we can see that this model over-estimates the non-overlap (108%), captures 27% percent of the overlap of the data and captures 0.01% percent of the

extreme overlap points.

One last interesting coefficient to look at is the correlation coefficient described in section 3.5 by equation (3.8). We calculated the correlation from the data and the correlation of the simulated data from the two-vehicle model for the 44 combinations of vehicles and averaged this over the 44 vehicles. In the following table the results can be found. The Danaher two-vehicle model does not capture the correlation very well.

	average correlation coefficient
Data	0.0773
Danaher model	0.1793

Table 3.3: Correlation coefficient for the data and simulated data from the Danaher model.

Given the results described above, the Danaher model does not seem to be a good representation of the vehicle-exposure data, since the model has a modest performance in capturing overlap and fails to capture extreme overlap. The Copula model might be a valid alternative with a better fit, which will be discussed in the next chapter.

4

Model 2 - Copula

4.1. Introduction

Since we are looking at overlap between vehicles, it would be useful to look at dependencies between vehicles and their mutually reached individuals. Therefore we will look at copula models for the comparison to the Danaher model. Copula models are multivariate distribution functions that describe the dependence between random variables. You can view them as functions that "couple" univariate distribution functions, which is exactly what we are looking for!

In this chapter we will first explain some theory about copulas, then fit and compare different copula models to our data and show the results obtained from simulating from these copula models.

4.2. Copula theory

The information in this section is based on the book *An Introduction to Copulas* by Nelsen (2006) [14] and *Coping with Copulas* by Schmidt (2006) [11]. Further details and proofs may be found therein.

Definition 4.2.1. A d-dimensional copula $C : [0,1]^d \longrightarrow [0,1]$ is a function which is a cumulative distribution function with uniform marginals. We notate a copula by $C(\mathbf{u}) = C(u_1, ..., u_d)$.

Since C is a distribution function, this leads to the following properties:

- As cdfs are increasing, $C(u_1, ..., u_d)$ is increasing in each component u_i.
- The marginal in component i is obtained by setting $u_j = 1$ for all $j \neq i$ and it must be uniformly distributed: $C(1, ..., 1, u_i, 1, ..., 1) = u_i$

The goal of copulas is to disentangle marginals and dependence structure.

Due to the following (and most important) theorem by Sklar (1959) we can entangle a copula and marginal distribution such that we end up with a proper multivariate distribution function.

Theorem 4.2.2. Sklar (1959). Consider a d-dimensional cdf F with marginals $F_1, ..., F_d$. There exists a copula C such that

$$F(x_1, ...x_d) = C(F_1(x_1), ..., F_d(x_d)) \tag{4.1}$$

for all x_i in [-∞, ∞], $i = 1, ..., d$. If F_i is continuous for all $i = 1, ..., d$ then C is unique; otherwise C is uniquely determined only on $Ran\, F_1 \times \cdots \times Ran\, F_d$, where $Ran\, F_i$ denotes the range of the cdf F_i.

On the other hand, consider a copula C and univariate cdfs $F_1, ..., F_d$. Then F as defined in (4.1) is a multivariate cdf with marginals $F_1, ..., F_d$.

4.2.1. Correlation

We know that copulas describe dependencies between random variables, but we are also able to connect a number to these dependencies using dependence measures. These measures, called correlation estimators, are very important since they can be represented in a form related to copulas, which give a possible way of fitting copulas to data for some parametric families. The two most commonly used estimators are Kendall's tau and Spearman's rho. We will only look at Kendall's tau for the correlation estimator and the fitting of copulas to our data. More information on Spearman's rho can be found in *Coping with Copulas* by Schmidt (2006) [11].

Kendall's tau correlation coefficient is a non-parametric measure which is defined using the concept of concordance. Consider two random variables X_1 and X_2, which have a joint distribution for random vector (X_1, X_2). Now consider a second independently but identically jointly distributed random vector (Y_1, Y_2). A pair is called **concordant** if $(X_1 - Y_1) \cdot (X_2 - Y_2) > 0$ and **discordant** when $(X_1 - Y_1) \cdot (X_2 - Y_2) < 0$. This results in the following definition of Kendall's tau:

Definition 4.2.3. We define Kendall's tau by:

$$\tau_K(X_1, X_2) = P((X_1 - Y_1) \cdot (X_2 - Y_2) > 0) - P((X_1 - Y_1) \cdot (X_2 - Y_2) < 0)$$

We can see that if both probabilities are the same, i.e. the same probability for concordant and discordant pairs, we obtain $\tau_K = 0$ which means the random vectors are independent.

The interpretation of the Kendall's tau coefficient is the same as other correlation coefficients: the measure has values in [-1,1], where 0 means two variables are independent, -1 means a negative correlation, +1 means a positive correlation. In our problem the Kendall's tau value for vehicles X_1 and X_2 can be interpreted as follows:

- $\tau_K(X_1, X_2) = 0$: seeing vehicle X_1 does not give any information on the chance of then also seeing vehicle X_2.
- $\tau_K(X_1, X_2) \in [-1, 0)$: seeing vehicle X_1 gives a smaller chance of also seeing vehicle X_2.
- $\tau_K(X_1, X_2) \in (0, 1]$: seeing vehicle X_1 gives a higher chance of also seeing vehicle X_2.

We can also estimate Kendall's Tau from our data. In general this can be done in in the following way: given a matrix of data containing n observations with columns $X = (X_1, ..., X_n)$ and $Y = (Y_1, ..., Y_n)$, let $Con = \#concordant\ pairs$ and $Dis = \#discordant\ pairs$. Then Kendall's tau can be estimated by

$$\tau_K = \frac{Con - Dis}{n(n-1)/2}$$

We will see more on Kendall's tau later in this chapter.

4.3. Important copulas

In this subsection we will present the mostly used copulas, which we will later fit to our data. We will present them using their scatterplots corresponding to the values 0.3 and 0.7 of Kendall's tau, to show the difference between a weak and strong correlation. The sample size for all scatterplots is n = 3612, which is the number of individuals in our selected data. We will also give the relationship between the parameters of the copula family and the Kendall's tau coefficient. The theory given in this subsection is based on *Modelling finite mixture joint distribution* by Bakker (2020) [10], *Coping with Copulas* by Schmidt (2006) [11] and *Relationship Between Kendall's tau Correlation and Mutual Information* by Ghalibaf (2020) [12]. A more detailed explanation can be found therein. The two most recognised classes of copulas are Elliptical copulas and Archimedean copulas. We will now give some theory on copulas within these classes.

4.3.1. Elliptical copulas

Gaussian/Normal copula. The Normal/Gaussian copula with parameter $\rho \in (-1, 1)$ is:

$$C_\rho(u_1, u_2) = \Phi_\Sigma(\Phi^{-1}(u_1), \Phi^{-1}(u_2))$$

where Σ is the 2×2 matrix with 1 on the diagonal and ρ otherwise, i.e., the correlation matrix. Φ denotes the cdf of the standard normal distribution and Φ_Σ is the cdf for the bivariate normal distribution with zero mean and covariance matrix Σ. This representation is equivalent to:

$$C_\rho(u_1, u_2) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi\sqrt{1-\rho^2}} exp\left(-\frac{s_1^2 - 2\rho s_1 s_2 + s_2^2}{2(1-\rho^2)}\right) ds_1 ds_2$$

The relationship between the parameter of the normal copula ρ and Kendall's tau is $\rho = sin\left(\frac{\pi}{2}\tau_k\right)$.

Example scatterplots for parameter ρ equal to 0.4540 and 0.8910 corresponding to Kendall's tau coefficients of 0.3 and 0.7 respectively can be found in the following figure:

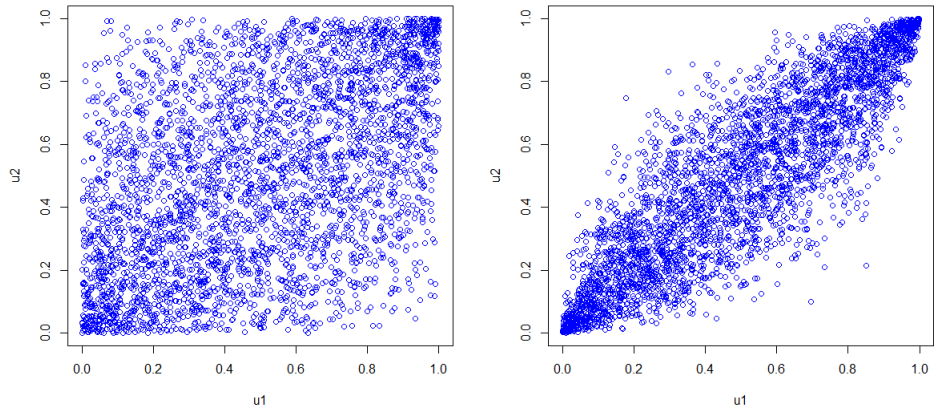


Figure 4.1: Scatter plots from normal copulas with parameters corresponding to Kendall's tau of 0.3 (left) and 0.7 (right).

We can see that the Gaussian copula is symmetric, with a larger concentration of points in the lower left and upper right corner. When the Kendall tau is higher, the points become more concentrated around the diagonal and there are almost no points in the upper left or lower right corner.

Student t-copula. The Student t-copula with parameters $\rho \in (-1, 1)$ and $\upsilon \geq 2$ is

$$C_{\upsilon,\Sigma}(u_1, u_2) = t_{\upsilon,\Sigma}(t_\upsilon^{-1}(u_1), t_\upsilon^{-1}(u_2))$$

where Σ is the same matrix as for the Gaussian copula, but with parameter ρ for student's t, υ is the degrees of freedom of the copula, t_υ is the cdf of the univariate student's t-distribution and $t_{\upsilon,\Sigma}$ the cdf of the bivariate student's t-distribution. This representation is equivalent to:

$$C_{\rho,\upsilon}(u_1, u_2) = \int_{-\infty}^{t_\upsilon^{-1}(u_1)} \int_{-\infty}^{t_\upsilon^{-1}(u_2)} \frac{1}{2\pi(1-\rho^2)^{\frac{1}{2}}} \left(1 + \frac{s_1^2 - 2\rho s_1 s_2 + s_2^2}{\upsilon(1-\rho^2)}\right)^{-\frac{\upsilon+2}{2}} ds_1 ds_2$$

The relationship between the parameter of the student t-copula ρ and Kendall's tau is $\rho = sin\left(\frac{\pi}{2}\tau_k\right)$.

Example scatterplots for parameter ρ equal to 0.4540 and 0.8910 corresponding to Kendall's tau coefficients of 0.3 and 0.7 respectively and parameter υ equal to 3 can be found in the following figure:

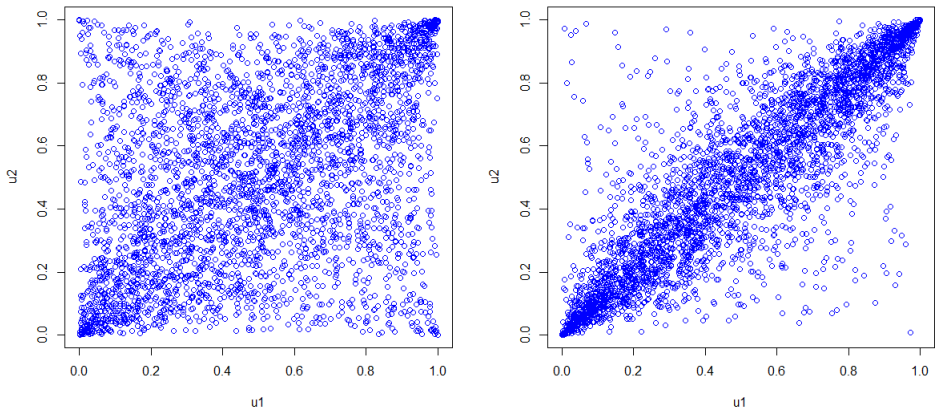


Figure 4.2: Scatter plots from student t-copulas with parameters corresponding to Kendall's tau of 0.3 (left) and 0.7 (right) and 3 degrees of freedom.

Comparing the Gaussian copula and Student t-copula we can see that the Gaussian is more centered around the diagonal, while the Student t has more spread with higher densities in the four corners. This shows that the Student t has higher tail dependence than the Gaussian. Tail dependence is the probability that u_1 reaches large values given that u_2 reaches large values (i.e. the corners are more dense). The two variables behave more closely in the corners with respect to the center in the Student t-copula than in the Gaussian copula.

4.3.2. Archimedean copulas

The general form of bivariate Archimedean copulas is defined as follows:

$$C(u_1, u_2) = \phi^{-1}(\phi(u_1) + \phi(u_2))$$

where ϕ is called the **generator function** of the copula. ϕ must be a decreasing function mapping [0,1] into [0,∞]. Every Archimedean copula has its own generator function.

Theorem 4.2.4. Consider a continuous and strictly decreasing function ϕ : [0,1] $\longrightarrow$ [0,∞] with $\phi(1) = 0$. Then

$$C(u_1, u_2) = \begin{cases} \phi^{-1}(\phi(u_1) + \phi(u_2)) & if\ \phi(u_1) + \phi(u_2) \leq \phi(0) \\ 0 & otherwise \end{cases}$$

is a copula, if and only if ϕ is convex.

Thus $C(u_1, u_2)$ is a copula if it satisfies Theorem 4.2.5. Using the definition and theorem one is able to generate quite a number of copulas. We will now give the most important examples of these copulas.

Clayton copula. The Clayton copula with parameter $\alpha \in [-1, \infty) \setminus 0$ is:

$$C_\alpha(u_1, u_2) = max([u_1^{-\alpha} + u_2^{-\alpha} - 1]^{-\frac{1}{\alpha}}, 0)$$

with generator function

$$\phi_\alpha(t) = \frac{1}{\alpha}(t^{-\alpha} - 1)$$

The relationship between the parameter α of the Clayton copula Kendall's tau is $\alpha = \frac{2\tau_k}{1-\tau_k}$.

Example scatterplots for parameter α equal to 0.8571 and 4.6667 corresponding to Kendall's tau coefficients of 0.3 and 0.7 respectively can be found in the following figure:

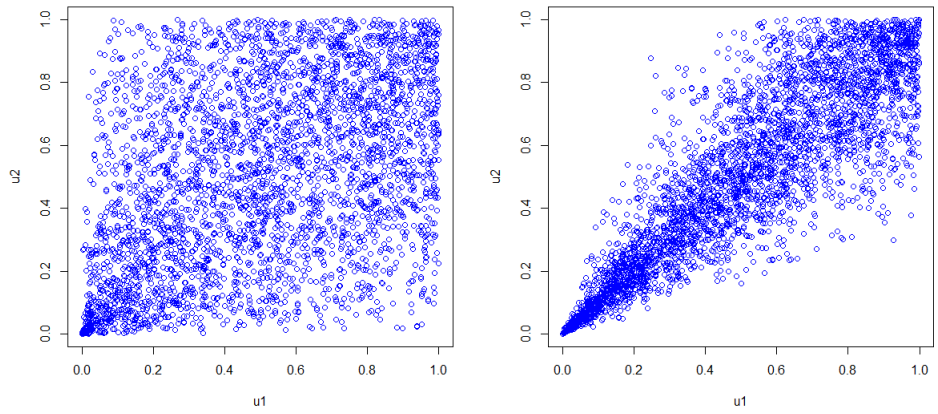


Figure 4.3: Scatter plots from Clayton copulas with parameters corresponding to Kendall's tau of 0.3 (left) and 0.7 (right).

We can see that the Clayton copula is very asymmetrical, with a very strong lower tail dependence.

Gumbel copula. The Gumbel copula with parameter $\alpha \in [1, \infty]$ is:

$$C_\alpha(u_1, u_2) = exp\left(-\left[(-ln(u_1))^\alpha + (-ln(u_2))^\alpha\right]^{\frac{1}{\alpha}}\right)$$

with generator function

$$\phi(t) = (-ln(t))^\alpha$$

The relationship between the parameter α of the Gumbel copula and Kendall's tau is $\alpha = \frac{1}{1-\tau_k}$.

Example scatterplots for parameter α equal to 1.4286 and 3.3333 corresponding to Kendall's tau coefficients of 0.3 and 0.7 respectively can be found in the following figure:

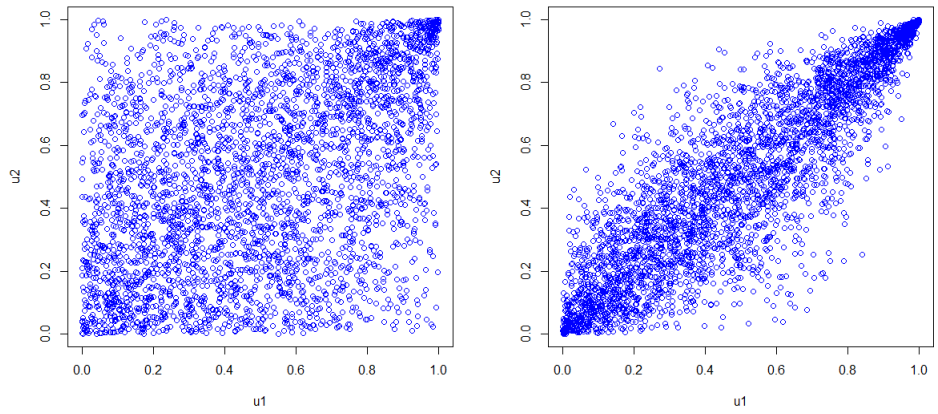


Figure 4.4: Scatter plots from Gumbel copulas with parameters corresponding to Kendall's tau of 0.3 (left) and 0.7 (right).

We can see that the Gumbel copula behaves asymmetrically, with a stronger upper tail dependence.

Frank copula. The Frank copula with parameter $\alpha \in (-\infty, \infty) \setminus 0$ is:

$$C_\alpha(u_1, u_2) = -\frac{1}{\alpha} ln\left(1 + \frac{(e^{-\alpha u_1} - 1)(e^{-\alpha u_2} - 1)}{e^{-\alpha} - 1}\right)$$

with generator function

$$\phi_\alpha(t) = -ln\left(\frac{e^{-\alpha t}-1}{e^{-\alpha}-1}\right)$$

The relationship between the parameter α of the Frank copula and Kendall's tau is $\tau_k = 1 + \frac{4}{\alpha}(D_1(\alpha)-1)$. Where D_k is the Debye function defined as:

$$D_k(\alpha) = \frac{k}{\alpha^k}\int_0^\alpha \frac{t^k}{e^t-1}dt$$

Example scatterplots for parameter α equal to 2.9332 and 11.4362 corresponding to Kendall's tau coefficients of 0.3 and 0.7 respectively can be found in the following figure:

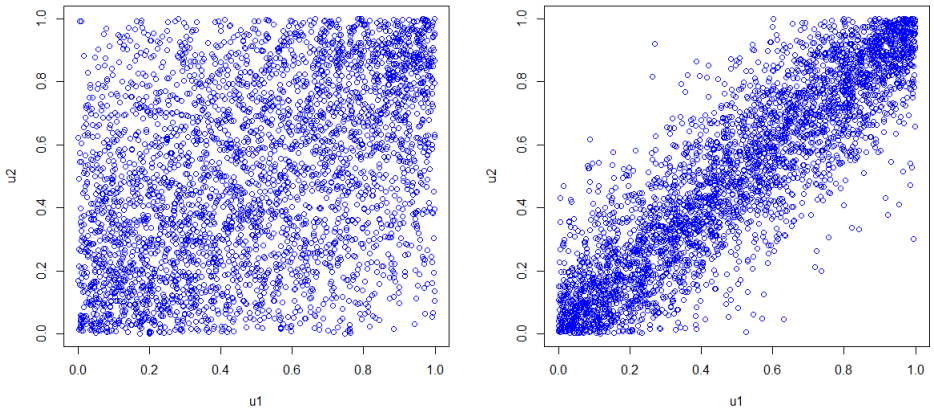


Figure 4.5: Scatter plots from Frank copulas with parameters corresponding to Kendall's tau of 0.3 (left) and 0.7 (right).

From the scatterplots we can see that there is no upper or lower tail dependence since the points seem evenly spread in the diagonal.

Let us compare the three types of Archimedean copulas described above. In figure 4.6 we displayed the three scatterplots of the copulas which have parameters corresponding to a Kendall tau coefficient of 0.7.

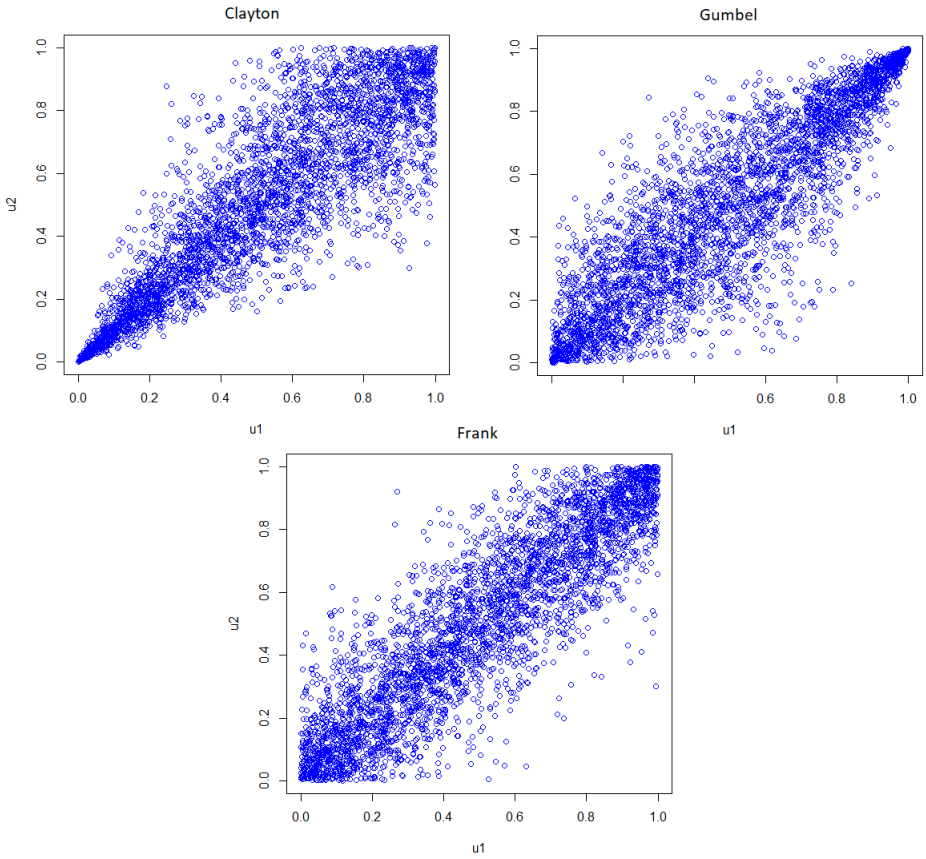


Figure 4.6: Scatterplots of Clayton (upper left), Gumbel (upper right) and Frank (lower) with parameters corresponding to Kendall tau of 0.7.

We can see that the Clayton copula has a very strong lower tail dependence in comparison to the other two copulas. Gumbel has stronger upper tail dependence than the other two. Only the Frank copula is symmetrical.

For every copula we can also calculate tail dependence coefficients, which is a number between 0 and 1 that describes the upper and lower tail dependence, giving an insight in dependencies of regions of the domain and whether there is a stronger dependency either in the upper or lower tail. Intuitively this can be explained by looking at a scatterplot. For the Clayton copula in figure 4.6, we can see that the plot is dense around the lower left corner, i.e. a strong lower-left tail dependence, meaning that the variables become **more** correlated for smaller values. In our problem this means that for the combination of vehicles X_1 and X_2 if an individual sees vehicle X_1 a small value of times, then the probability of seeing vehicle X_2 a small value of times becomes higher due to the high positive correlation. We will not go further into this theory, but if you are interested in these coefficients I advise you to read *Modelling finite mixture joint distributions* by Bakker (2020) [10].

4.4. Copula fitting

To choose which copulas we are going to fit and test to our data we first take a look at our data itself. In the following figure we can see the data points (left) and the data transformed to pseudo-observations (right). Pseudo-observations are transformed data points which are transformed to [0,1] scale using the empirical distribution or another parametric distribution function. Due to Sklar's theorem (4.1) this is needed for our copula models.

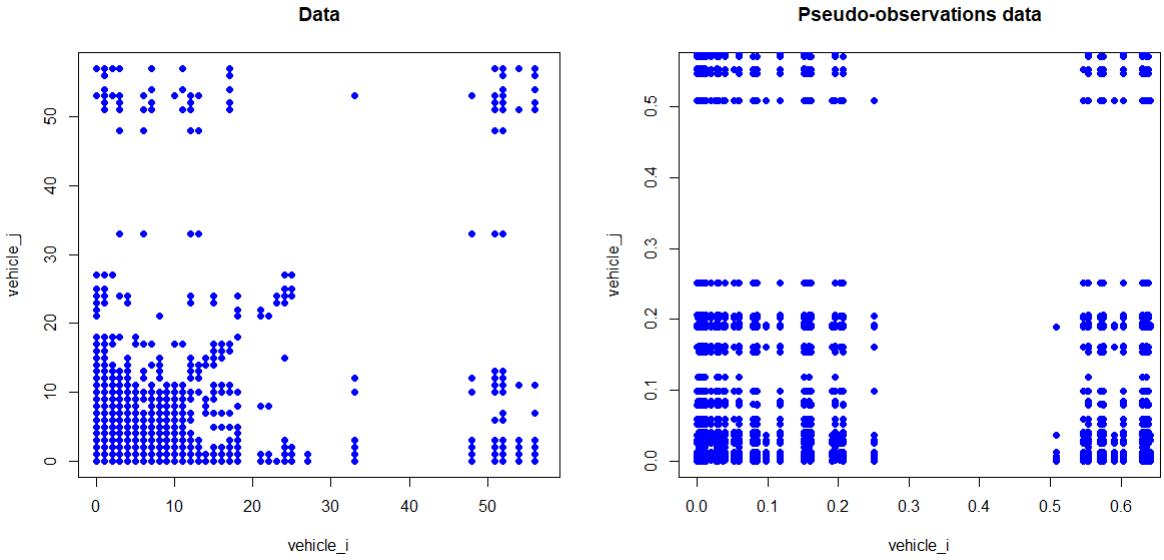


Figure 4.7: The data points (left) and the transformed pseudo-observations from data points (right).

There is a concentration of the datapoints in the lower-left corner of the graph, with fewer datapoints in the other corners. This is an expected result, since the probability of seeing any vehicle multiple times is low (top left and bottom right corners), and even less likely to see both vehicles many times (top right corner). The diagonal represents the overlap between vehicles, where both vehicles have been seen a number of times. Fitting the copulas to this data was challenging due to the presence of ties (repeating values), possible alternatives will be discussed in chapter 8 and 8.

We can compare the pseudo-observations in figure 4.7 to the examples of copulas given in section 4.3. In the following figure we see the difference between the pseudo-observations from the data and the different types of copulas. All parameters correspond to a Kendall tau of 0.3, which is about the average Kendall tau of the data.

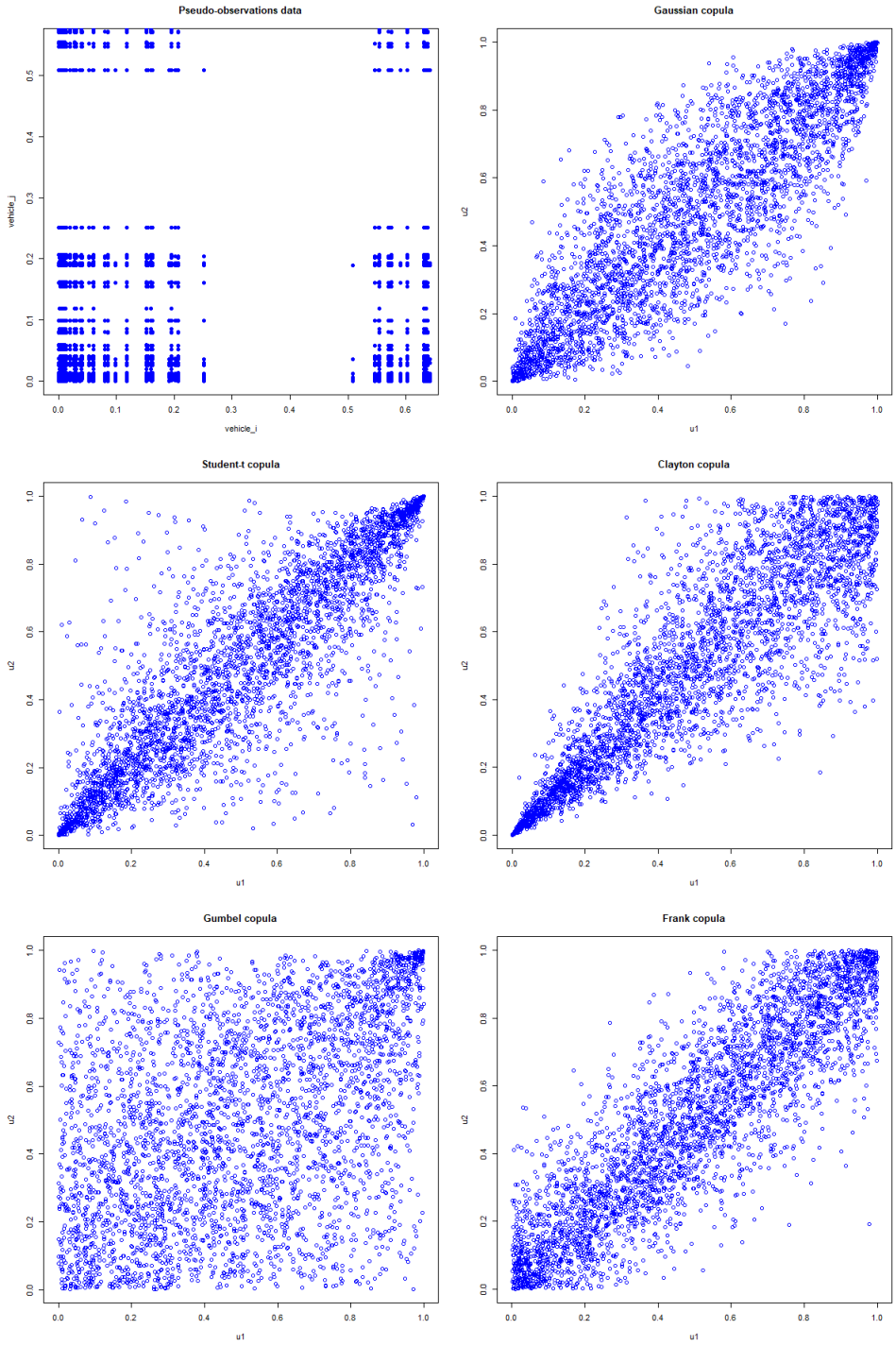


Figure 4.8: Pseudo-observations and copula simulations on the uniform interval. Parameters copula correspond to Kendall tau of 0.3. Subfigures represent the data (upper left), Gaussian (upper right), Student-t (middle left), Clayton (middle right), Gumbel (lower left) and Frank copula (lower right).

Looking at the pseudo-observations of the data we can see that it is very dense in lower left tail, which is why a copula with a strong lower left tail dependence could be a good fit to our data. The other three corners, especially the upper left and lower right, also have some density. Therefore, the copula should also have some points in these corners. Looking at this figure 4.8, we would expect the Clayton or Student-T to be quite a reasonable fit.

Something else that stands out is the difference between the data and copulas. This is due to the discrete-

ness of our data, which is why the fit of any copula does not seem to be an optimal fit. A solution to this problem could not be explored due to the time limitations, but will be discussed in chapter 6 and 8 for future studies.

We can check the performance of the different models by fitting copulas and checking the goodness-of-fit. We have two ways for fitting and simulating from copulas (in R, *copula* package [8]):

- Parametric: fit parametric univariate marginals using the *fitdist()* function, then fit copulas using the *fitCopula()* function to get the parameters of the copula family. Using these parameters and the parametric marginals we can specify a joint distribution using the function *mvdc()*. Then sample from this joint distribution using the function *rmvdc()*.

- Non-parametric: we transform the data to pseudo-observations using the *pobs()* function, then fit a copula using the *fitCopula()* function and get the parameters of the copula family. Then sample from the multivariate copula using the *rCopula()* function to obtain [0,1] sample observations, then transform these back the original data-type using the *quantile()* function.

We need to transform the data to pseudo-observations due to Sklar's Theorem 4.1, which says that a copula is a function of uniform marginals. Therefore we need transform the margins of our dataset to standard uniform margins, to obtain copula data. In this thesis, we will focus on the second way of fitting the copulas. The implemented functions used for the fitting of the copulas can be found in Appendix A.2.1.

Using the functions described above we will fit bivariate copulas to all 253 pairwise combinations of vehicles that we have in our selected data. We will compare the fit of the following copulas to all the combinations of vehicles: Gaussian, Student-t, Gumbel, Clayton and Frank. In addition to these we will also fit a different type of copula: a copula that fits a copula family to every pairwise combination (which can be different for every combination) using the *BiCopSelect()* function in R, and combine all these. We call this copula the **selectCopula**. This function has a large set of copula families to choose from, which can be found in the *VineCopula* package [9].

All estimations for the parameters for the fit of the copula functions are done by using maximum likelihood, but due to our data, the optimization method used in the function that is used in R is different for some copulas. However, this does not significantly interfere with the outcomes of our model.

4.5. Results

We again use the selected regional data from London city center as described in chapter 2 to fit and test the different copula models to our data. The data consists of 253 pairwise combinations that we will consider in the model. The implementation of the model is done in R and the used functions can be found in Appendix A.2.1.

4.5.1. Simulating copulas

In the figure below you can find the simulated data points from the selectCopula, Gaussian, Student-t, Clayton, Gumbel and Frank copula for all pairwise combinations of vehicles. We can compare these simulated datapoints to the real datapoints.

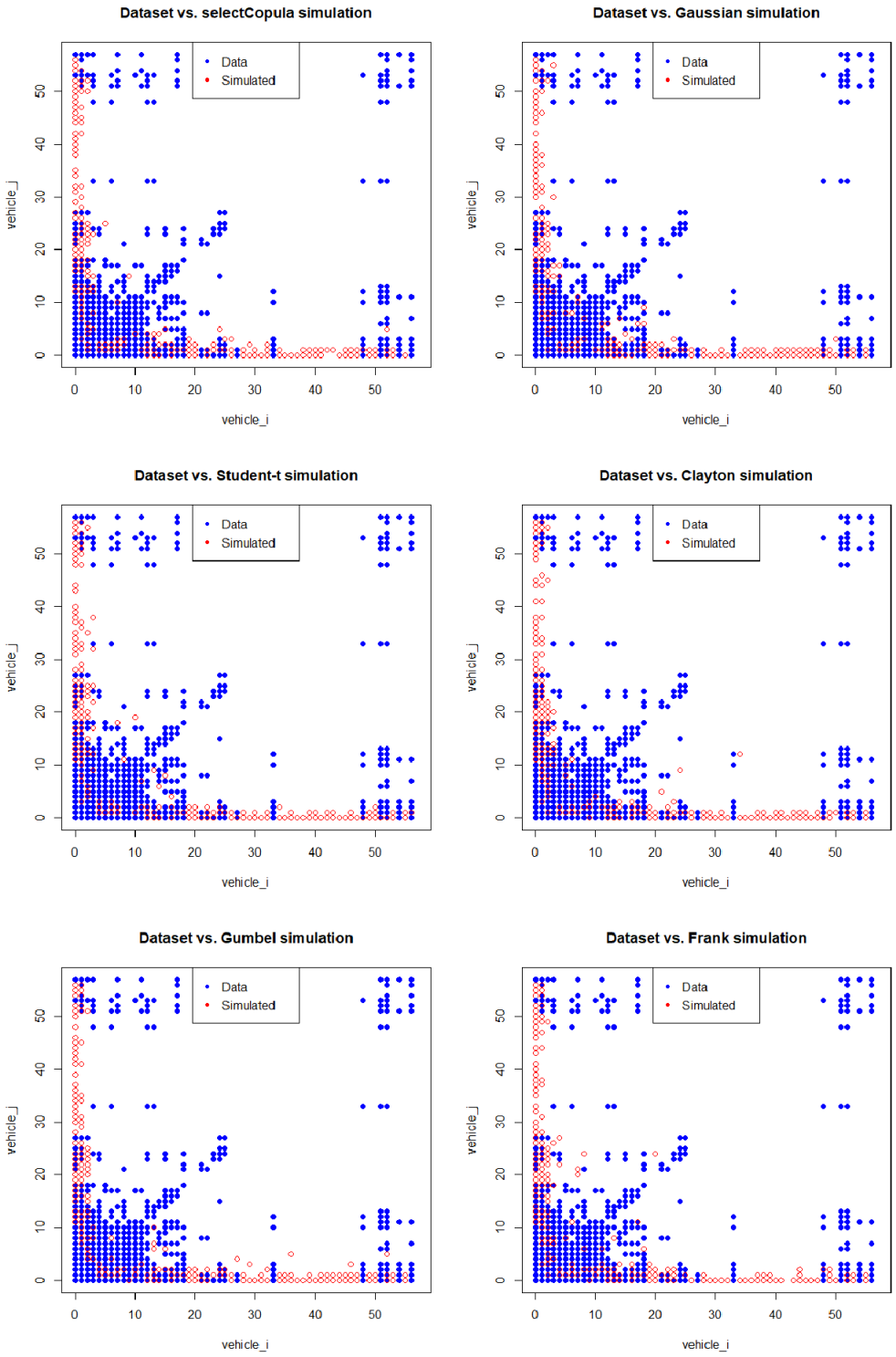


Figure 4.9: Simulated data points from selectCopula (upper left), Gaussian (upper right), Student-t (middle left), Clayton (middle right), Gumbel (lower left) and Frank (lower right).

We can see that all copulas look alike: the simulations are close to the x and y axis. This means that they do not really capture overlap. We can also see that all copulas do simulate extreme values, but only for combinations where one of the two vehicles has a high frequency and the other has a small frequency. The Student-t, Clayton and Frank do have some extreme overlap combinations, in comparison to the others.

4.5.2. Goodness-of-Fit

To test which model is the best fit to our data we would like to use the *gofCopula()* function that empirically compares the empirical copula with a parametric estimate of the copula under the null hypothesis. But due to the long running time of this function (~ 1 hour for one combination of vehicles) we cannot perform this analysis. Therefore we will look at the average log-likelihood of the fit of the copulas, which are given by the *fitcopula()* function for each fit. The average is taken over all 253 combinations of vehicles. The higher the log-likelihood, the better the model performs. In this way, we can compare the copula models to each other and see which model fits best. Additionally we can use the same criteria as for the Danaher model explained in section 3.4.

Log-likelihood. We calculate the log-likelihood for the fit of every combination of vehicles, then take the average of these likelihoods to obtain the average log-likelihood of the fit. In the following table you can find the results of the different copula models.

	selectCopula	Gaussian	Student-t	Clayton	Gumbel	Frank
average log-likelihood	2416.18	616.35	1730.64	3030.13	646.70	940.40

Table 4.1: Average log-likelihoods for copula models.

The higher the log-likelihood, the better fit of the model to our data. Based on the log-likelihood criteria the Clayton model performs best, followed by the selectCopula and Student-t copula. However, we would expect the selectCopula to perform best since this functions fits the "best" copula to the combination of vehicles we are looking at. A possible explanation for this difference could be due to the optimization method used for the optimization of maximum likelihood, which will be discussed further in chapter 8.

Capturing overlap. The following table presents three vehicle combination examples and their count for non-overlap, overlap and extreme overlap. You can also see the summed counts for the overlap categories over all the combinations of vehicles. Lastly, the percentage of the capture with respect to the data is shown for the three overlap categories.

vehicle combinations	data	selectCopula	Gaussian	Student-t	Clayton	Gumbel	Frank	type overlap
(1,2)	3277	3390	3301	3270	3322	3300	3271	
(2,3)	3467	3587	3610	3490	3654	3434	3695	non-overlap
(1,3)	3114	3178	3334	3264	3532	3312	3236	
(1,2)	355	222	311	342	290	312	341	
(2,3)	145	25	2	122	42	178	83	overlap
(1,3)	498	434	278	348	80	300	376	
(1,2)	11	3	2	2	1	3	0	
(2,3)	0	0	0	1	1	1	0	extreme overlap
(1,3)	8	1	2	1	3	2	1	
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
(0,0), (i,0), (0,j) : i,j > 0	812766	860108	860011	860559	860099	860104	860285	non-overlap
(i,j) : i,j > 0	101070	53728	53825	53277	53737	53732	53551	overlap
(i,j) : i,j > 4	2102	29	30	35	36	38	15	extreme overlap
	100%	106%	106%	106%	106%	106%	106%	non-overlap
captured percentage w.r.t. data	100%	53%	53%	53%	53%	53%	53%	overlap
	100%	1.38%	1.43%	1.67%	1.71%	1.81%	1.57%	extreme overlap

Table 4.2: An overview of the captured non-overlap, overlap and extreme overlap for the copula models.

All models perform about the same, therefore the capturing overlap criterion does not give us much information on the difference in performance of the copula models. However, the Gumbel, Clayton and Student-t have a small better performance in capturing extreme overlap.

We can also compare the correlation coefficient of the different combinations of vehicles of the data to the correlation coefficient of combinations of the simulated data. We calculate this coefficient using Kendall tau.

We will compare the average Kendall tau coefficient of all vehicle combinations to the Kendall tau coefficient of the simulated data for the vehicle combinations. In the following table you can see the results:

	Data	selectCopula	Gaussian	Student-t	Clayton	Gumbel	Frank
average Kendall tau	0.1870	0.1858	0.1868	0.1866	0.1872	0.2452	0.1877
difference from data	0	0.0012	0.0002	0.0004	0.0002	0.0482	0.0007

Table 4.3: Average Kendall tau for data and copula models. The average is calculated by looking at the Kendall tau correlation between two vehicles for all 253 pairwise vehicle combinations.

We want the Kendall tau values of the copula models to be as close as possible to the Kendall tau of the data. We can see that the Gaussian and Clayton copulas have the closest Kendall tau to the one from the data, but all have quite small differences.

In conclusion, looking at the several criteria described above, not one model distinctively performs better than the others. In the following chapter we will compare the Danaher model from chapter 3 and the copula models.

5

Model comparison

5.1. Introduction

In this chapter we will compare the two-vehicle Danaher model from chapter 3 to the copula models from chapter 4. Based on two different measures we will select the best fit of the models: the figures of the simulated points versus the actual dataset, and the capture of the overlap from counting the non-overlap, overlap and extreme overlap. All results are explained more detailed in the chapters on the Danaher model (3) and Copula model (4).

5.2. Comparison

Simulations. In the following figure the simulations from the Danaher model (44 vehicle combinations) and the Clayton copula model (all 253 combinations) can be found. Since all copula models from chapter 4 perform almost equally, we only present the Clayton simulations for a better overview for the comparison to the Danaher model. The data is also plotted.

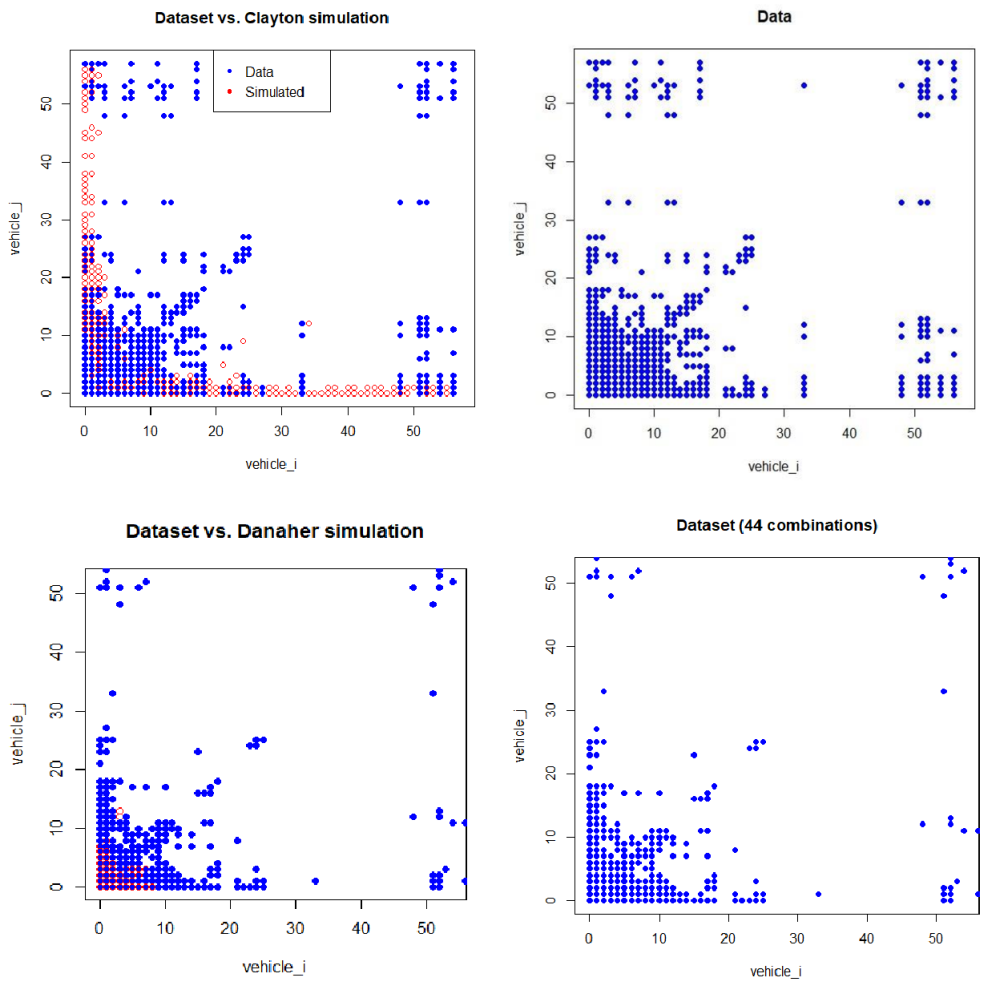


Figure 5.1: Simulations from the Clayton copula model (upper left) and Danaher model (lower left). On the right side the data points are plotted for comparison. The Danaher model is simulated for 44 combinations, the copula model for 253 combinations.

Comparing the two left subfigures to the right we can see that the Danaher model simulates the lower right corner better, especially small values, while the copula model does not. The copula model does simulate the overlap points better than the Danaher model, but overall both models do not fit the data optimal.

Capturing overlap. We value the capture of the overlap since the overlap most is very interesting for advertisements. If two vehicle have a large overlap, most of the times it would be better to choose one of the two vehicles for an advertisement, since they reach about the same people, and spent the budget on another spot that reaches different individuals. In the following table the results for both the Danaher model and the different copula models are presented for the percentage of captured non-overlap, overlap and extreme overlap compared to the observed data. The Danaher percentages are based on the 44 combinations of vehicles and the copula percentages are based on the 253 combinations of vehicles.

	vehicle combinations	data	Danaher	selectCopula	Gaussian	Student-t	Clayton	Gumbel	Frank	type overlap
captured	(0,0), (i,0), (0,j) : i,j > 0	100%	108%	106%	106%	106%	106%	106%	106%	non-overlap
percentage	(i,j) : i,j > 0	100%	27%	53%	53%	53%	53%	53%	53%	overlap
w.r.t. data	(i,j) : i,j > 4	100%	1.49%	1.38%	1.43%	1.67%	1.71%	1.81%	1.57%	extreme overlap

Table 5.1: The captured percentage w.r.t. data for the non-overlap, overlap and extreme overlap data points between the data and simulated data from the Danaher and different copula models.

It can be seen that the copula models perform modestly better in capturing the non-overlap, overlap and

extreme overlap when comparing it to the Danaher model.

In conclusion, a Copula model is a better fit to our data than the Danaher model. However, both models fit the data far from perfect. Alternatives to the Danaher and Copula models will be discussed in chapter 8, along with suggestions for further research.

6

Adding distance

6.1. Introduction

In chapter 1 of this thesis we already explained the intuitive relationship between overlap and distance: two vehicles placed closely together would expectedly show a higher overlap. In chapter 2 we presented the following figure:

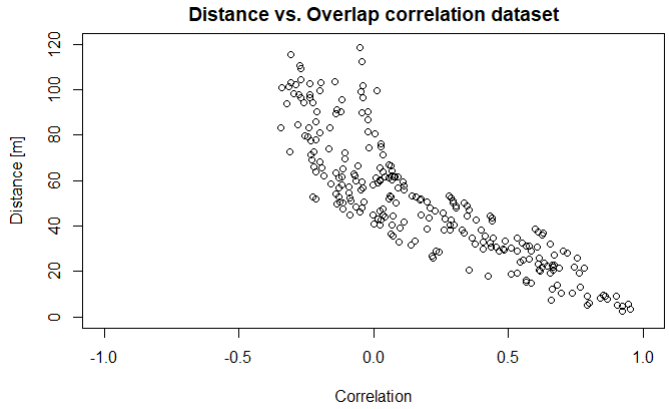


Figure 6.1: The Kendall tau correlation between two vehicles with respect to the distance between the two vehicles.

where the Kendall tau correlation between all the 253 combination of pairwise vehicles is plotted against the distance in meters. This figure shows there is a higher correlation for small distances, when the distance becomes larger the correlation coefficient becomes more centered around zero. As visible from figure 6.1 there appears to be a relationship between the distance and overlap between vehicles.

In this chapter we will research this relationship further. Looking at the Danaher model and Copula model, the Copula model offers an extension to 3-dimensions (or more), so we can fit a 3-dimensional copula to model the dependency between the exposures of individuals to two vehicles and the distance between these vehicles. This extension is not yet established for the Danaher model, therefore we will focus this chapter on the fitting of 3-dimensional copulas. We will compare the following 3-dimensional copulas: Gaussian, Student-t, Clayton, Gumbel and Frank.

6.1.1. Modifying data

In the Danaher and Copula model described in chapters 3 and 4 respectively we fitted the models on the data of a combination of two vehicles, then doing this for all combinations of vehicles and putting all these fits together. In this way, we can model the overlap between two specified vehicles, taking only into account the exposures of individuals to these two vehicles. If we want to add the variable distance, we have to "modify" the data we are fitting the models to, since the distance is constant when looking at just the data from two

specified vehicles, i.e., the distance between the vehicles stays the same so the third variable would be a constant which would not give us any information. Therefore we need to find a way to fit the models to the exposures of the individuals to all the different combinations of vehicles, where the distance variable will not be constant. We do this by making a 3-column matrix, where the first two columns represent the exposures of the individuals to all the combinations of vehicles, and the third column is the distance between the vehicles of which the individual is exposed to. In table 6.1 examples of the data used in chapter 3 and 4 is compared to the data used in this chapter. We added the column distance to the first subtable to make the comparison clearer, the real data used in chapter 3 and 4 is a two-dimensional matrix consisting of the first two columns.

	vehicle 1	vehicle 2	distance [m]
person 1	2	1	12
person 2	0	3	12
.	.	.	.
.	.	.	.
.	.	.	.
person 3612	1	1	12

(a) Example data used in Danaher (chapter 3) and Copula (chapter 4) models, with distance added for clearer comparison.

	vehicle i	vehicle j	distance [m]
person 1	2	1	12
.	.	.	.
.	.	.	.
.	.	.	.
person 3612	1	1	12
person 1	0	0	25
.	.	.	.
.	.	.	.
.	.	.	.
person 3612	10	8	25
.	.	.	.
.	.	.	.
.	.	.	.

(b) Example data used when adding variable distance in chapter 6.

Table 6.1: Example data for comparison between the data used in chapter 3/4 and for adding variable distance.

In subtable (a) the distance is constant since we are only looking at the exposures of individuals to the vehicles 1 and 2. Using this data we can fit the Danaher model and the Copula model to model the frequencies and overlap for these two vehicles, then repeat this process for all the other combinations of vehicles and combine the results. In subtable (b) all the exposures of the individuals to all the combinations of vehicles is combined, and it can be seen that the distance is now varying for every combination of vehicles. In this chapter we will use the format of the example data in table 6.1 (a).

The newly obtained data can be show in a 3-dimensional scatterplot. Figure 6.2 presents this scatterplot from different angles.

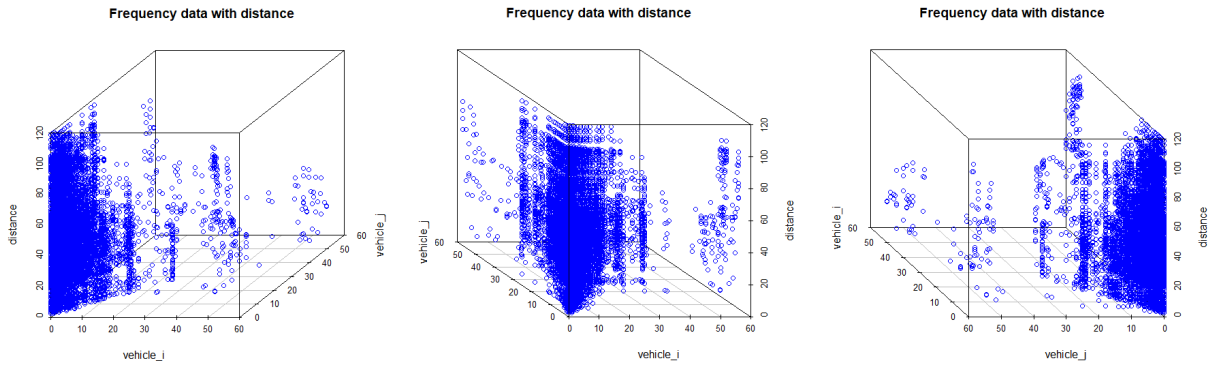


Figure 6.2: 3-dimensional data representing the exposures of individuals to all combinations of vehicles and the distance between those vehicles. These three plots are from the same data but presented from different angles.

The points are mostly centered around the distance axis for small values of exposures, since these combinations of exposures are present for all combinations of vehicles and therefore for all possible distances. The interesting phenomena that can be seen in the scatterplots is in the overlap diagonal of the $vehicle_i$ and $vehicle_j$ axis. There is downward slope from high distances to lower distances in the diagonal of the frequencies plane (the plane of $vehicle_i$ and $vehicle_j$). This means that when the overlap becomes larger when the distance becomes smaller.

6.2. Fitting 3-dimensional copulas

We will start by fitting the 3-dimensional copulas using the same method used in chapter 4 for fitting the 2-dimensional copulas. This means transforming our data to pseudo-observations and fitting different copulas to estimate the parameters of the copula families. Next we will simulate from these copulas to obtain sample observations and compare these to the observed data.

In the following figure the data and pseudo-data are plotted:

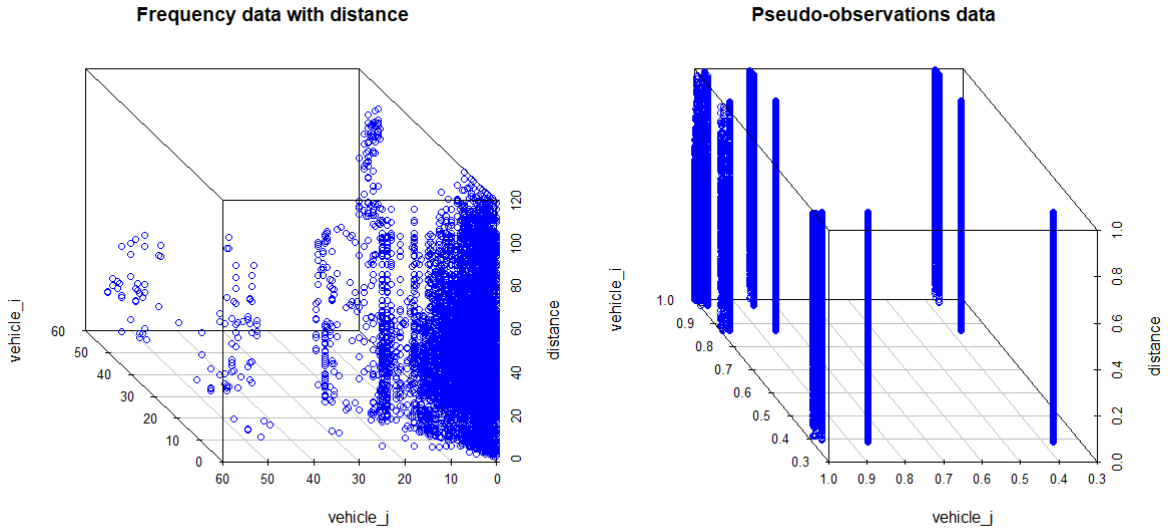


Figure 6.3: Data and transformed pseudo-data.

There are a few high peaks at the corners of the plot. This is due to the nature of our data and the large amount of ties in the exposures of individuals. We fit the copulas using these pseudo-observations and simulate to compare the result to the observed data. The results can be found in figure 6.4.

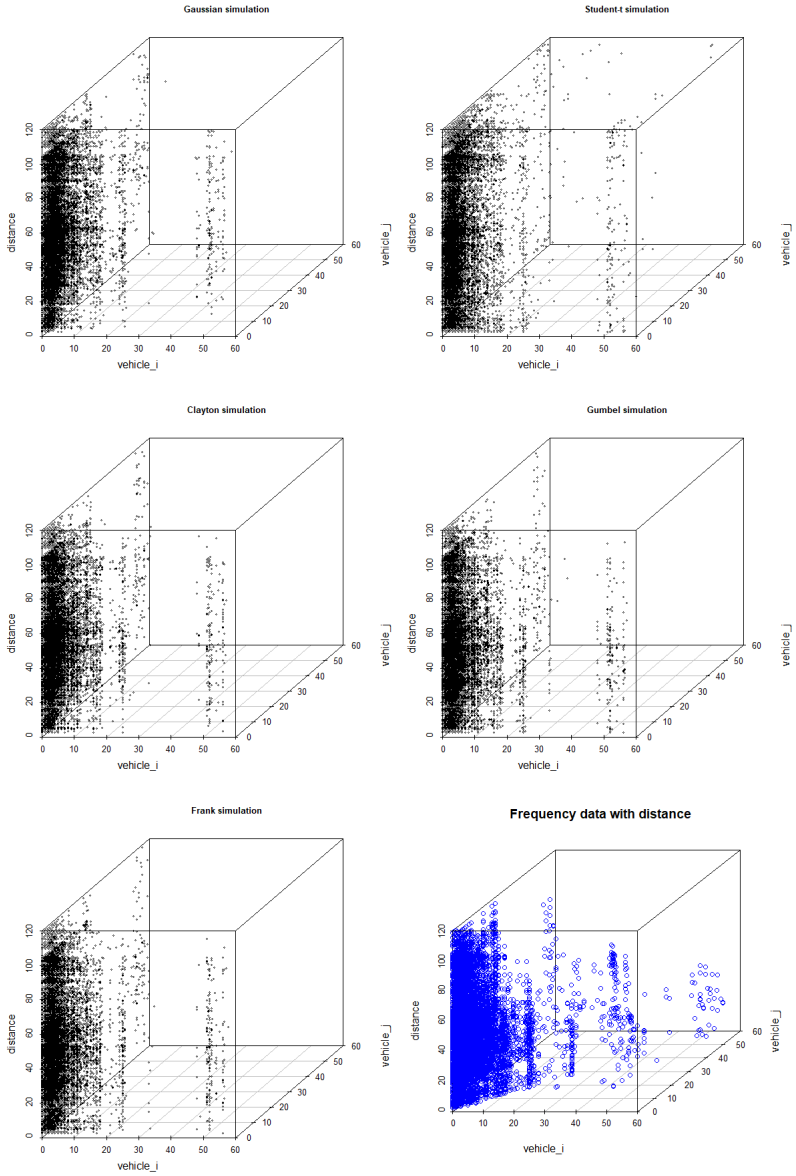


Figure 6.4: Copula simulations fitted by using transformed pseudo-observations. For comparison the 3-dimensional scatterplot of the data is also added (lower right).

We can see that the copulas fit the frequencies and distances of the smaller frequency values good, but for larger frequency values the distance is not representing the distance in the observed data. We cannot clearly see how the simulations capture the overlap. Therefore we plot the top-views of the copulas, which are presented in the following figure:

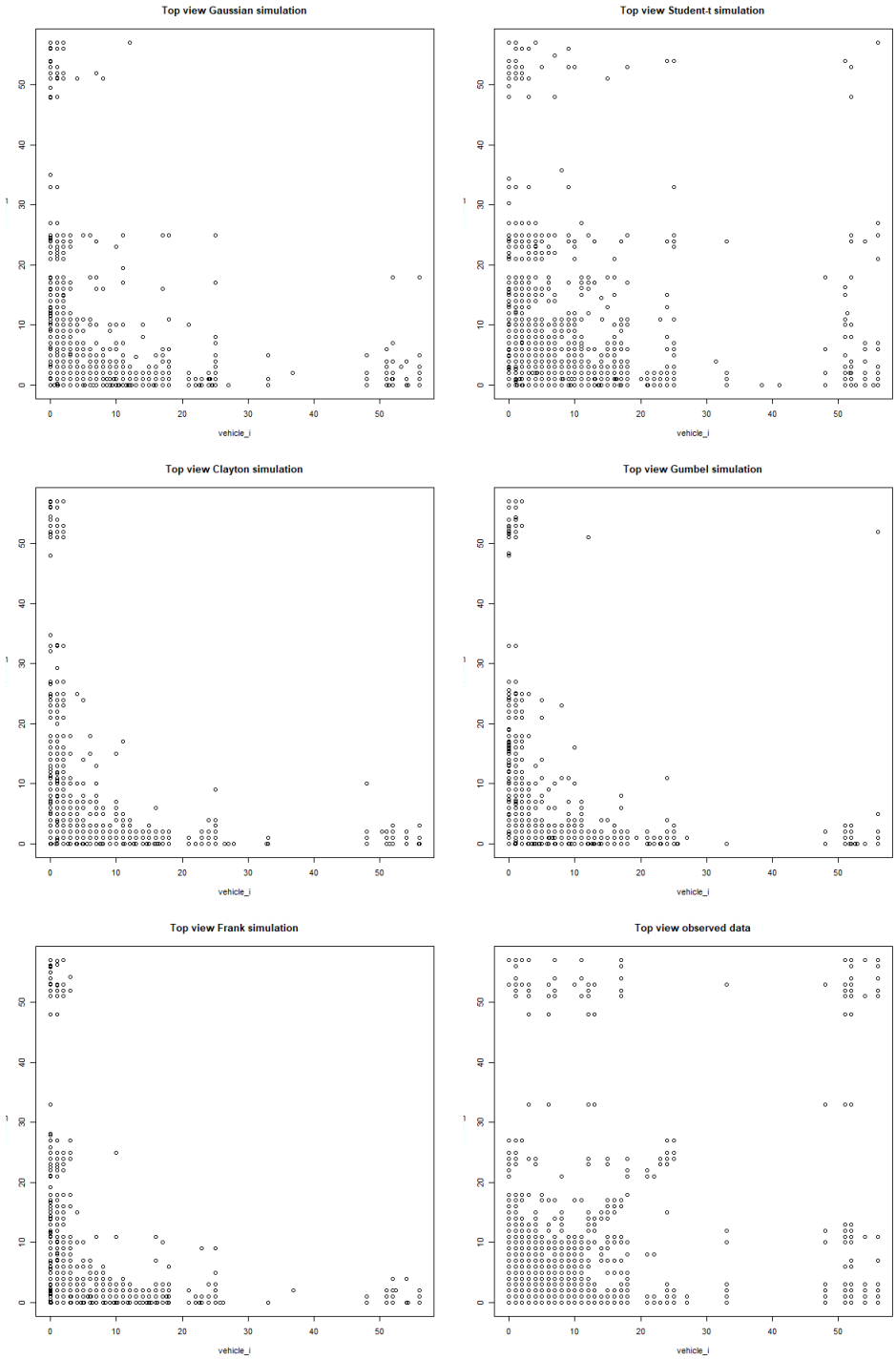


Figure 6.5: Top-view different copula models. The top-view of the observed data is added for comparison.

The Student-t captures the overlap best and does this better than any previously presented model in this thesis. This is a very particular result. The Student-t copula simulates the overlap excellent and even simulates some extreme overlap points. We want to try to improve the capturing of the extreme overlap even more. We will try two different methods to improve this capture:

- Fitting continuous parametric marginals to the data to change the piece-wise linear cdf to a smooth curve. This curve ensures that the data does not contain ties anymore. Then fit the copulas using the *fitCopula* function in R. We construct a multivariate copula using the *mvdc()* function , then simulate the copula models using the *rMvdc()* function in R.

- Creating some small noise around the data, called jittering. This is a function in R (*jitter()*) that creates a noise around the data points with a certain factor for the amount of noise added. Using this function you transform the data to one without ties. Then fit copulas and simulate from these copulas using the same procedure as before.

Both options create a certain change in the data, which also means they lead to biased estimations of the parameters of the copula families. However, this could also lead to a better fit of the copulas to our data due to the removal of ties. In the following paragraphs we will explore both options and show the results.

Using parametric marginals. We start the procedure of this method by fitting univariate parametric marginals to the data. We do this for each of the three columns in our data, the frequencies of individuals to the combinations of vehicles and the distance. We need to do this for all three variables since the *mvdc()* functions needs three marginals specified for the 3-dimensions of our problem.

The function $model\_select()$ in R fits different parametric families and returns the family with fitted parameters for the model with the highest log-likelihood. We fitted the following distributions: Laplace, Logistic, Normal, Exponential, Gamma, Inverse Gamma, Inverse Gaussian, Log-logistic, Log-normal, Rayleigh, Weibull, Log-gamma, Pareto, Beta, Kumar, Negative binomial. In the following table the distributions with the highest log-likelihood are presented along with their estimated parameters:

variable	distribution	log-likelihood	parameters
frequency 1	exponential	119391	rate = 3.098
frequency 2	exponential	88958	rate = 2.998
distance	weibull	-4260873	shape = 1.98, scale = 57.46

Table 6.2: Fitted distributions and estimated parameters for every variable in our data using $model\_select()$ function in R.

The next step in the procedure is simulating points from these distributions and transforming the newly simulated data to pseudo-observations. We do this using the *rexp()* and *rweibull()* functions in R for the exponential and weibull distributions. We sample n = 50000 observations. The *pobs()* function is again used for transforming the samples to pseudo-observations. Next we fit the different copulas to the pseudo-observations using the *fitcopula()* function, then combine the fitted 3D copula and the univariate marginals using the *mvdc()* function. Lastly we simulate from our 3D model using the *rMvdc()* function.

We did this for the Gaussian, Student-t, Clayton, Gumbel and Frank copula families. Unfortunately we could not fit a Clayton and Frank copula due to the exponential behaviour in the zeros which created infinite values in the optimization process. In the following figure the simulations for the Gaussian, Student-t and Gumbel model can be found:

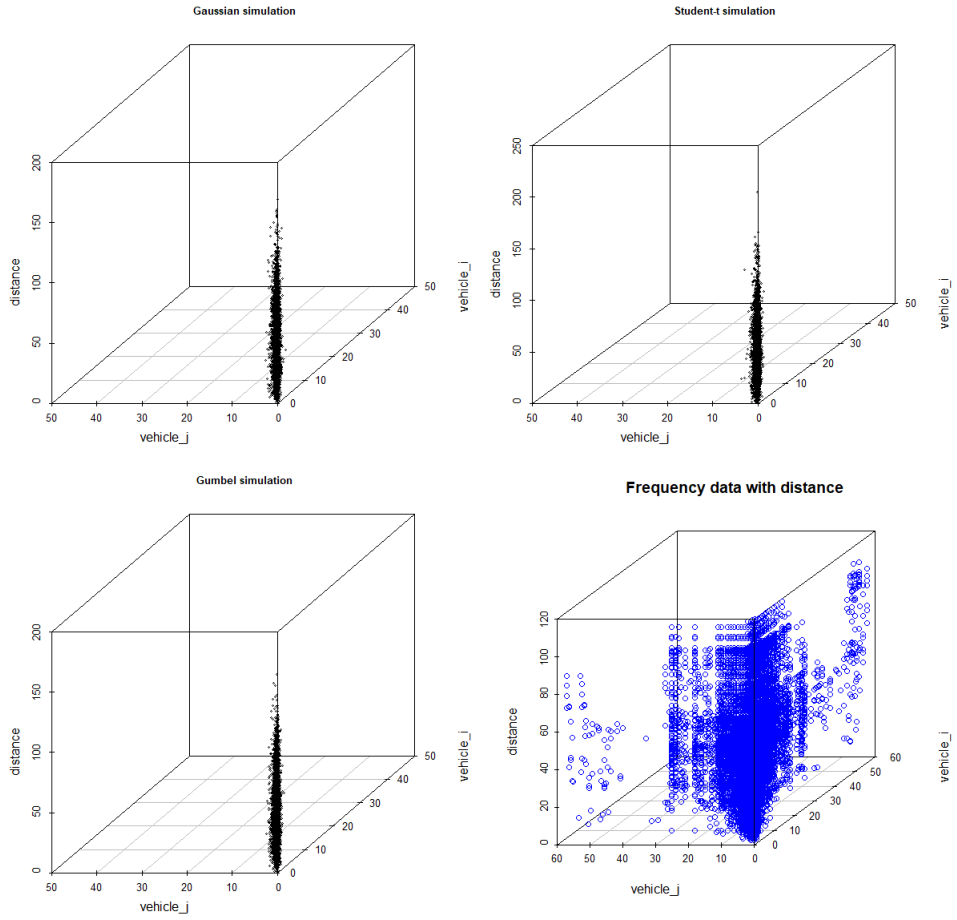


Figure 6.6: Simulations from the Gaussian (upper left), Student-t (upper right) and Gumbel (lower left) copula models. The observed data (lower right) is plotted for comparison.

It can be seen that the simulated points for all copula models are centered around the lower values of frequencies. The models fit the lower values of frequencies adequate, but they do not fit the higher values or extreme overlap at all. A main reason for this outcome is the nature of the data. It seems that there are still too many zeros in our selected data, thereby parametric families are fitted to support this many zeros, which results in a decrease in likeliness for high frequencies to occur with the use of parametric families, especially for the exponential distribution.

There could be a solution to get rid of this exponential behaviour at the zeros. We could use a different distribution for the choice of marginals for the frequencies in the fitting process of making our 3-dimensional copula. In the Danaher model we fitted the Negative Binomial Distribution (NBD) to the univariate data of the different vehicles. In section 3.6.1 we concluded that the NBD is a reasonable fit to our data, and therefore we will also try the NBD to see if using this distribution removes the exponential behaviour at the zeros and is a better representation of our data.

We repeat the exact same process as described above but now using the NBD for the marginals of the frequency columns in our data. When fitting the different copulas there did not occur an optimization error thus we were able to simulate from all copulas. The results can be found in figure 6.7.

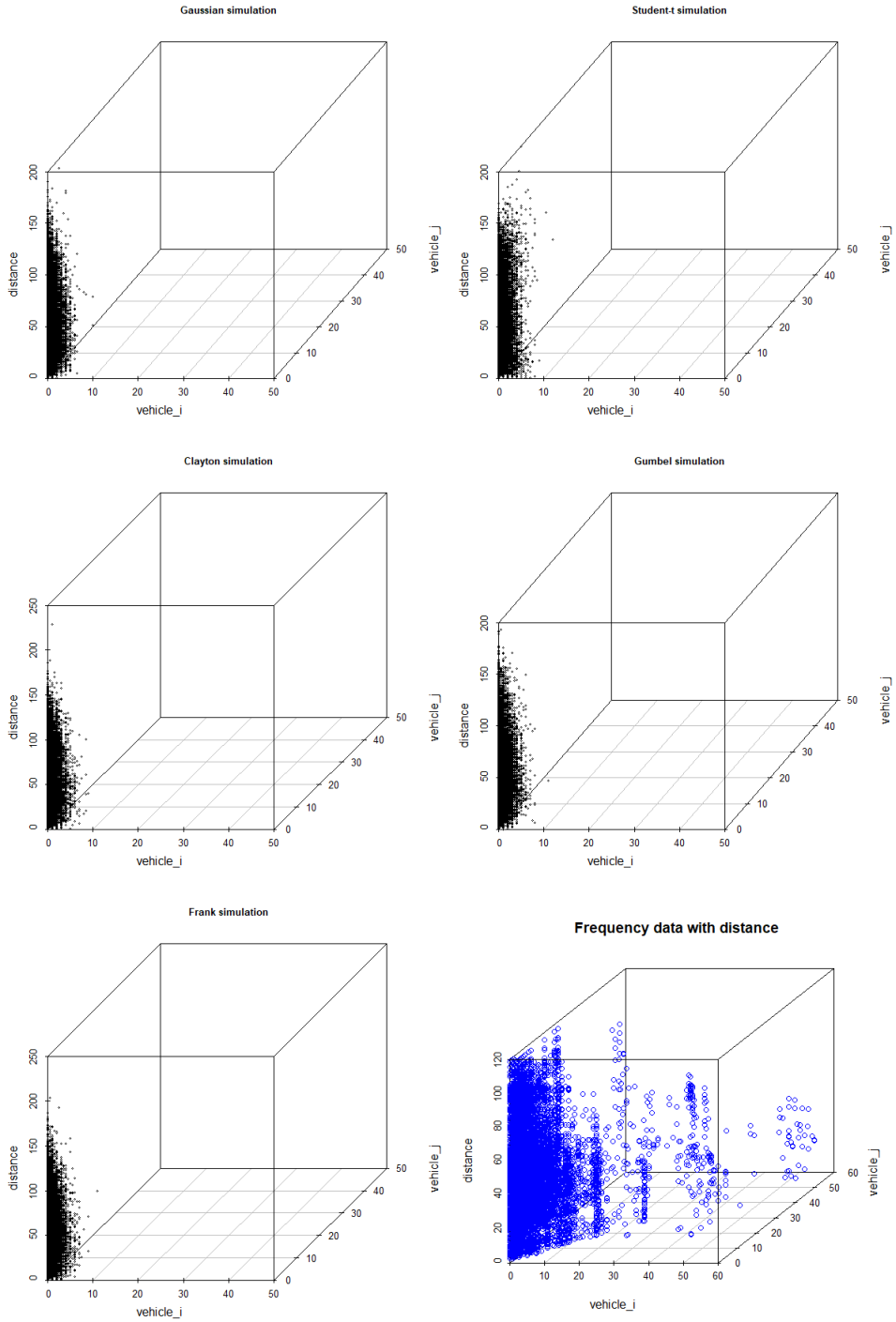


Figure 6.7: Simulations using the NBD as marginals for univariate frequency data of the Gaussian (upper left), Student-t (upper right), Clayton (middle left), Gumbel (middle right) and Frank (lower left) copulas. The data (lower right) is also plotted for comparison.

The simulated points are a little less dense around the small frequencies than for the use of exponential marginals. Still the copulas do not simulate any high frequencies or extreme overlap. They do fit the data better than for the use of the exponential marginals.

In conclusion, the method for fitting parametric marginals does not improve the performance of our 3-dimensional copula model. We will explain the second method in the following paragraph.

Jittering. We can either use the *jitter()* function on the data or the transformed pseudo-observations of the data. We tried both and for the data it created too much noise, since when transforming the data to pseudo-observations it magnified the noise. When we jittered on the already transformed pseudo-observations of the data we could compare different factors to see which factor to use for fitting the copulas. The factor can be interpreted as the amount of noise added to the data. In figure 6.8 the jittering of the pseudo-observations with factors equal to 0.05, 0.1, 0.3, 0.7, 1, 2, 3 and 5 are presented.

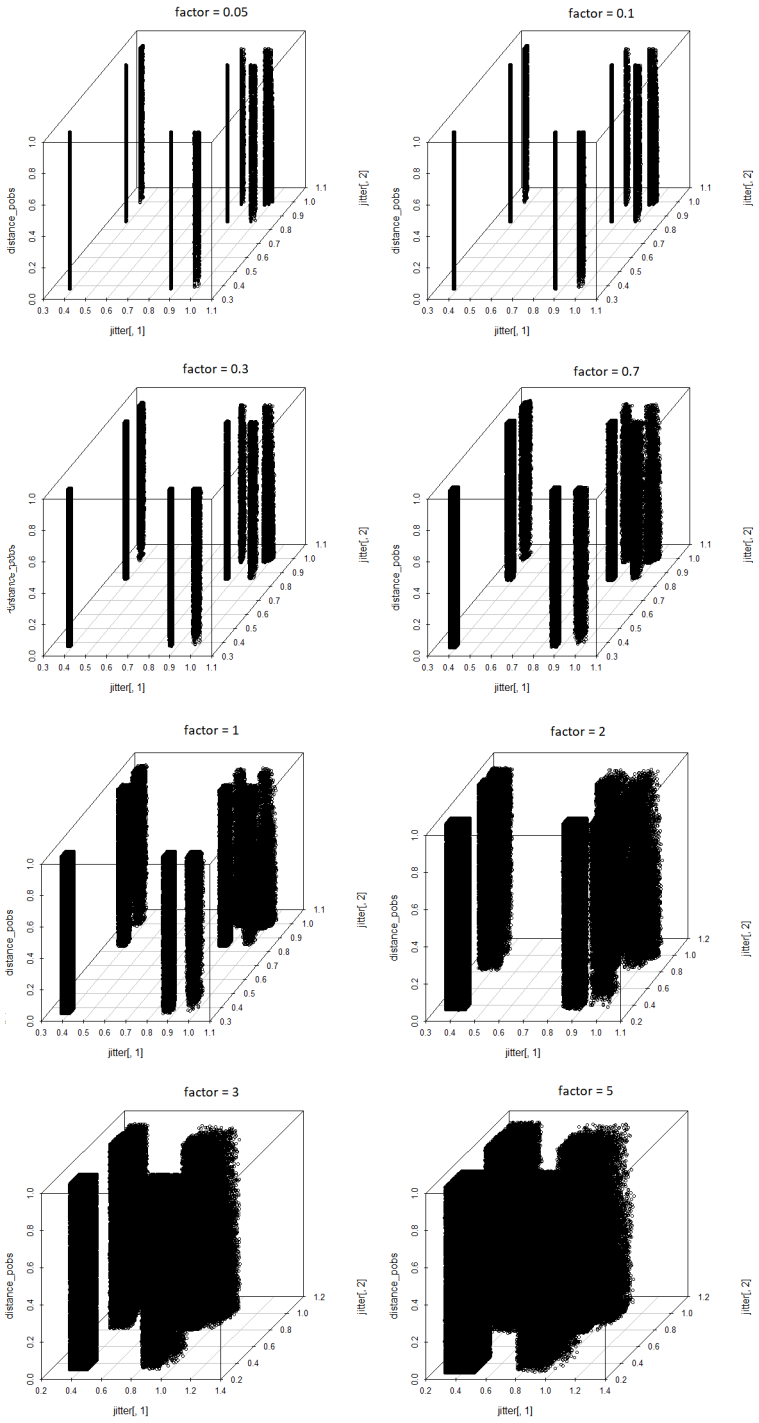


Figure 6.8: Jittered pseudo-observations data with factors equal to 0.05, 0.1, 0.3, 0.7, 1, 2, 3, and 5 from left to right, upper to lower plots respectively.

The higher the factor, the more noise added. To not create too much noise but still modify the data enough to remove ties we choose the factor argument equal to 1. Next we fit the different copulas and simulate from these copulas, the obtained results are observations in [0,1] for every variable. Due to the high computation time of the fitting for over 900 thousand jittered data points (> 2 hours) we randomly choose 50 thousand jittered data points to fit the copulas. Lastly we transform these [0,1] observations back to the original data using the *quantile()* function. The results can be found in figure 6.9.

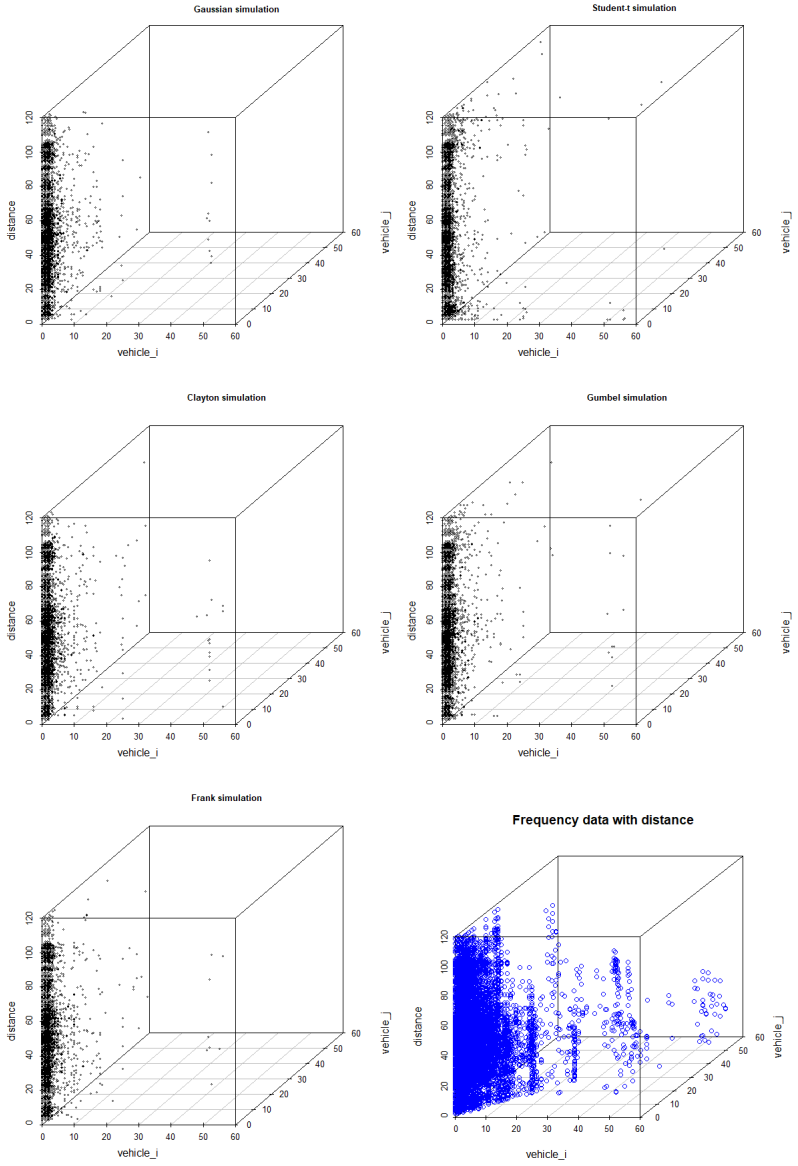


Figure 6.9: Simulations from different copula fits to jittered pseudo-observations. The data is added for comparison.

There is some improvement in the resemblance of the data when comparing it to the previous method, but still modest performance in capturing the overlap. In figure 6.10 the top-views of the simulations are plotted, along with the top-view of the data for comparison.

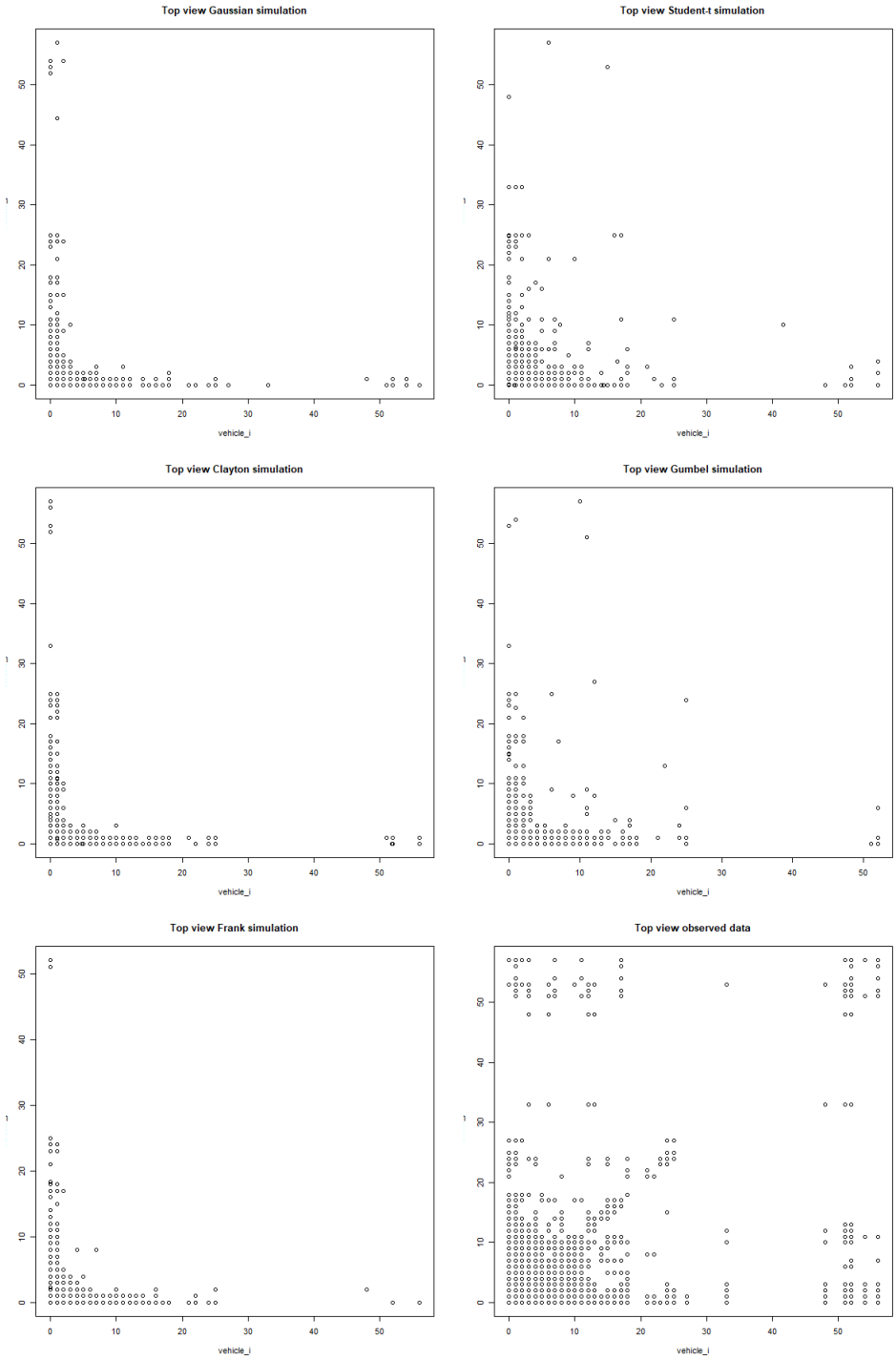


Figure 6.10: Top-view of simulated points from jittered copulas.

The jittering of the copulas does not improve the capturing of the overlap in comparison to the original simulated copulas in figure 6.5. Neither methods improved the fitting of the original 3-dimensional copula model.

6.2.1. Conclusion

In this section we saw that when we modify the data to a combination of the occurring exposures for all the different vehicle combinations, the simulations of the Student-t copula for the frequencies are a good resemblance of the observed exposure data. However, the 3-dimensional copula fails to simulate the relationship

of the distance and exposures between vehicles. Due to lack of time we will not investigate the relationship between distance and overlap further, this will be mentioned in the recommendations for further research in chapter 8.

6.3. Looking back at 2-dimensions

In chapter 4 we fitted copulas to the data of two specified vehicles, then did this for all combinations of vehicles and combined all the results. In this chapter we modified the data to first combine all the datapoints of all the different combinations of vehicles, then fitted copulas. This could be a better fit to our data since we can use more information for the fit of the copulas. In figure 6.5 we can see that the frequencies and overlap are very well resembling the observed data for the 3-dimensional Student-t copula. This result gives the impression that if we fit 2-dimensional copula models to the modified data, we could get a good resemblance of the data and a well performing model for the bivariate exposure distribution that we are researching in this thesis. We fitted the same copula families as used in the previous sections, the results are presented in the following figure:

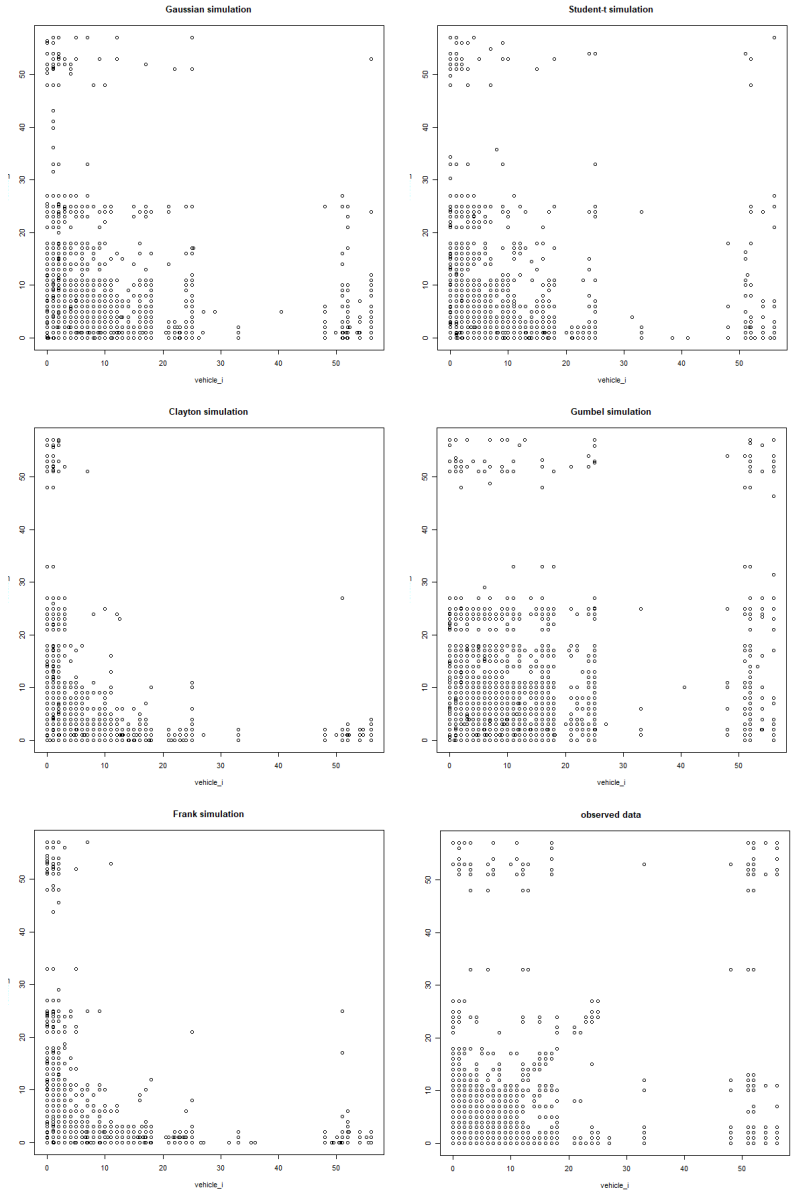


Figure 6.11: Fitted copulas to the modified 2-dimensional data. The observed 2-dimensional data is plotted for comparison.

The figure shows that the Gumbel and Student-t copula perform best in simulating non-overlap, overlap and extreme overlap. However, comparing the two copula models to the observed data we can see that the Gumbel copula overestimates the overlap, and the Student-t underestimates the extreme overlap. By just comparing the copulas to the data we conclude that the Student-t copula is the best fit to our data, but more thorough analysis can be done on the fit of the copulas. Due to lack of time we will not do this analysis but advise future researchers to do so.

This result shows that there is a Copula model that resembles our data, including the non-overlap, overlap and extreme overlap. This indicates that the Copula model is a good choice for the bivariate exposure distribution to model frequencies and overlap between vehicles.

As a final remark we would like to present the conditional probability distribution for the probability that there is a certain overlap between the exposures to two vehicles (variables X_1 and X_2), given the distance between those vehicles (variable X_3). This distribution [17] is given by:

$$\begin{aligned} F(X_1 \geq x_1, X_2 \geq x_2 | X_3 = x_3) &= 1 - F(X_1 \leq x_1, X_2 \leq x_2 | X_3 = x_3) \\ &= 1 - \frac{\partial C(F_1(x_1), F_2(x_2), F_3(x_3)}{\partial F_3(x_3)} \end{aligned} \tag{6.1}$$

where $F_1(x_1)$, $F_2(x_2)$ and $F_3(x_3)$ are the marginal CDF's corresponding to the variables. The conditional copula C describes the conditional dependence structure of (X_1, X_2) given $X_3 = x_3$. According to Sklar's theorem (4.2.2) this copula exists. Using equation (6.1) we could actually estimate the overlap in exposures of vehicles given the distance between them. Therefore, if we were able to find a copula that is a good fit to our 3-dimensional data, equation (6.1) can answer the question on the relationship between overlap and distance.

With this result we will finish this research. In the following chapters we will conclude and discuss the findings of this thesis.

7

Conclusion

The goal of this thesis was modelling bivariate exposure distributions for Out of Home advertising, to model the frequencies of individuals to media vehicles and modelling the overlap between vehicles. We compared two different models: the Danaher and Copula model. Both models had significant presence within marketing science and were worthwhile investigating.

The Danaher model was based on fitting a Negative Binomial Distribution to the univariate data of vehicles. The univariate marginals were used in a specialised version of the Sarmanov Bivariate Distribution to create a bivariate exposure distribution. The results from the Danaher model were not a good fit to our data, as they did not capture high frequencies or overlap.

The Copula model was based on fitting commonly used copulas using a non-parametric approach. The fitting and simulating from copulas created some challenges and the results were inadequate in capturing overlap. The main reason for this result is the choice of data for fitting the copulas.

In addition to the comparison of these two models we extended the two-variable problem to a 3-variable problem by adding a distance variable. This variable represents the distance between vehicles and we wanted to see if we could model the relationship between the overlap and distance between vehicles. To investigate this relationship we modified the data to be able to take distance into account. The copula models showed good performance in fitting small exposure according to distance, but had modest performance in modelling high exposures/overlap and distances. Lastly we used the modified data to look back at the 2-dimensional Copula model, for which we found a good fit to the data using the Student-t copula. This Copula model performed optimal out of the proposed models and represents the data well.

Combining all the obtained results we can conclude that the 2-dimensional Student-t copula is the best model for our exposure distribution, with a good performance in capturing non-overlap, overlap and extreme overlap. Secondly, we could not find a model that accurately models the relationship between distance and overlap, while taking into account the peculiar characteristics of the exposure data.

8

Discussion

Throughout the course of this research we ran into quite some difficulties and challenges. In this chapter we will discuss these complications and give suggestions for future research.

The main reason is the nature of the data. We carefully selected the data based on geographical location, trying to model the frequencies of individuals that locate themselves in this region. This resulted in selecting a number of vehicles in a region and taking into account the exposures of individuals that have seen at least one of the vehicles. The data can be presented in a matrix, where all the columns represent the vehicles and the rows represent the individuals. The values in the matrix correspond to the frequency of the individual in row j with respect to the vehicle in column i. This creates some zeros (when an individual in the selected data does not see a certain vehicle), but in comparison to using the whole dataset we keep the zeros to a minimum. Throughout this research the zeros were still a challenge, along with the repeating of the same values (ties), which make fitting models difficult. A possible solution for the amount of zeros could be to change the minimum amount of vehicles an individual has to see, this gets rid of the zeros from individuals that only see one vehicle. We did not try this in this research but this could be interesting for further research. Some solutions to the ties problem have been explored in chapter 6. The methods did not improve the resemblance of the simulations to the data and other ideas are interesting to investigate.

Another big challenge was of a more technical nature. The running time of the code was quite long for some parts of models (> 2 hours for some parts). This resulted in having to make the sample size of the models lower to be able to obtain results. For the Danaher model in chapter 3 we had to randomly pick 44 combinations of vehicles (out of the 253) to fit and sample from the model. For the 3-dimensional copula model in chapter 6 we had to pick 50.000 random datapoints (out of the over 900.000 datapoints) to fit the copula models. These alterations in sample size could affect the results and therefore make the results questionable.

While researching the one-vehicle model in the Danaher chapter 3 we fitted the univiarite NBD and did a statistical test to see if the NBD fits the data well. Due to lack of time we only tested this for four vehicles (out of 23). To make this result complete all vehicles should be tested. For the two-vehicle model we were not able to do a bivariate goodness-of-fit test. Instead we altered the test used in the Danaher paper [1] to see if the models capture overlap adequately. To statistically test the goodness-of-fit of the two-vehicle model a proper goodness-of-fit test should be used, but it can be quite a challenge to find this test.

While researching the Copula model in chapter 4 we used the same type of data that is used for the Danaher model: bivariate data of two specified vehicles. The copula model can be improved by using the same type of data used in the 3-dimensional copula model from chapter 6. In section (6.3) we already fitted some copula models by using the data from the 3-dimensional copula, but due to lack of time we could not do a thorough analysis. Investigating this further could lead to even better results of the Copula model.

The problems mentioned above give a good reason to investigate both the Danaher and Copula model further. In addition to the challenges we ran into during this research and improvements that are given above,

we have some ideas on what topics to investigate for future research and the possibilities for researchers in marketing science.

8.1. Recommendations

In this section we will itemize some recommendations for further research.

- We carefully selected data to fit and simulate from the different models presented in this thesis. A second step could be to upscale the number of vehicles or enlarge the region. You could also compare different regions to obtain information about the frequencies/overlap in different cities/regions.
- In the chapters about copulas we focused on the most commonly used copulas, but there are many more copula families to be thought of, and some copulas can also be rotated to better represent the data.
- For both the Danaher and Copula model we fitted the parameters using maximum likelihood. There are lots of other methods to fit parameters, which are worthwhile investigating.
- We did not find a proper defined relation between the distance and overlap in this thesis. This is still something very much of interest in marketing science and definitely an interesting research topic.
- In this thesis we only looked at the comparison of two vehicles and modelled bivariate exposure distributions. A very interesting research topic could be to look at multivariate exposure distributions where we compare more than two vehicles.
- If we are able to model the exposures of individuals and estimate the overlap, we could make an optimization problem for picking the vehicles to minimise overlap within a certain budget.

A

Appendix

A.1. Danaher model

A.1.1. Maximum Likelihood Estimation for parameter p of the univariate NBD

The Probability Mass Function of the Negative Binomial Distribution is defined as:

$$P(X_i = x_i \mid r_i, p_i) = \binom{x_i + r_i - 1}{x_i} p_i^{r_i} (1 - p_i)^{x_i} \quad x_i = 0, 1, 2, ...$$

Likelihood function:

$$\begin{aligned} L(p) &= \prod_{i=1}^{n} Pr(x_i|p) = \prod_{i=1}^{n} \binom{x_i + r + 1}{x_i} p^r (1-p)^{x_i} \\ &= \left(\prod_{i=1}^{n} \binom{x_i + r - 1}{x_i} \right) p^{nr} (1-p)^{\sum_{i=1}^{n} x_i} \end{aligned}$$

where x_i are the n observations of random variable X. To maximize this function we take the logarithm to get the log-likelihood, differentiate and set this to zero, then solve to obtain the maximum.

$$ln(L(p)) = ln\left(\prod_{i=1}^{n} \binom{x_i + r - 1}{x_i} \right) + nr * ln(p) + \left(\sum_{i=1}^{n} x_i \right) * ln(1-p)$$

$$Set \frac{\partial L}{\partial p} = 0 \Longrightarrow \frac{\partial L}{\partial p} = \frac{nr}{p} - \frac{\sum_{i=1}^{n} x_i}{1-p} = 0 \Longrightarrow r(1-p) = \overline{x}p \Longrightarrow p = \frac{r}{\overline{x} + r}$$

We can check that p is a maximum by taking differentiating again. This number is negative thus p is a maximum. Our estimation for p is: $\hat{p} = \frac{r}{\overline{x}+r}$.

A.1.2. Implemented functions in R

```
---
title: "Danaher model functions"
author: "Frederiek Backers"
---

```{r, libraries and data import}
library(MASS)
library(fitdistrplus)
library(readxl)
library(feather)
library(plot3D)
```
```

```
library(ggplot2)
library(copula)
library(scatterplot3d)
library(univariateML)
library(VineCopula)
library(sads)
library(discreteRV)
library(plyr)

data <- arrow::read_feather("C:\\Users\\Frederiek.Backers\\Gitlab\\Code\\
                             Rdata_region_Londonzoom.feather")
colnames(data) <- 0:(length(data)-1)
```

```{r, functions}
# Obtain vehicle data
vehicle_data <- function(vehicle_nr){
  result <- data[[vehicle_nr+1]]
}

# Parameter estimation
# size = r
est_r <- function(vehicle_nr){
  result <- fitdist(vehicle_data(vehicle_nr), "nbinom", method=)$estimate[1]
}

# mu = mu (= mean)
est_mu <- function(vehicle_nr){
  result <- fitdist(vehicle_data(vehicle_nr), "nbinom", method=)$estimate[2]
}

# parametrization a (paper)
est_a <- function(vehicle_nr){
  result <- est_r(vehicle_nr)/est_mu(vehicle_nr)
}

# w
est_w <- function(vehicle_nr_1, vehicle_nr_2){
  # Make count function for estimation of f_hat
  count_single <- function(vehicle_nr){
    count = 0
    for(person in 1:length(vehicle_data(vehicle_nr))){
      if(data[[vehicle_nr+1]][person] == 0){
        count = count + 1
      }
    }
    result <- count
  }
  f_hat_single <- function(vehicle_nr){
    result <- count_single(vehicle_nr)/length(vehicle_data(vehicle_nr))
  }
  # Make count function for estimation of joint f_hat
  count_double <- function(vehicle_nr_1, vehicle_nr_2){
    count = 0
    for (person in 1:length(vehicle_data(vehicle_nr_1))){
      # we could also do length of vehicle_nr_2 but these are equal
```

```
      if(data[[vehicle_nr_1+1]][person] == 0 &
         data[[vehicle_nr_2+1]][person] == 0){
        count = count + 1
      }
    }
    result <- count
  }
  f_hat_double <- function(vehicle_nr_1, vehicle_nr_2){
    result <- count_double(vehicle_nr_1,vehicle_nr_2)
    /length(vehicle_data(vehicle_nr_1))
  }
  result <- (f_hat_double(vehicle_nr_1,vehicle_nr_2)/
              (f_hat_single(vehicle_nr_1)*f_hat_single(vehicle_nr_2)) - 1)*
               (1/(phi(0,est_r(vehicle_nr_1),est_a(vehicle_nr_1))*
               phi(0,est_r(vehicle_nr_2),est_a(vehicle_nr_2))))
}

# Calculate joint distribution
joint_distribution <- function(vehicle_nr_1, vehicle_nr_2, x_i, x_j){
  X_i = c(0:max(vehicle_data(vehicle_nr_1)))
  X_j = c(0:max(vehicle_data(vehicle_nr_2)))
  est_P_i = dnbinom(X_i,size=est_r(vehicle_nr_1),mu=est_mu(vehicle_nr_1))
  est_P_j = dnbinom(X_j,size=est_r(vehicle_nr_2),mu=est_mu(vehicle_nr_2))
  return <- est_P_i[x_i+1]*est_P_j[x_j+1]*(1+est_w(vehicle_nr_1,vehicle_nr_2)*
              phi(x_i,est_r(vehicle_nr_1),est_a(vehicle_nr_1))*
              phi(x_j,est_r(vehicle_nr_2),est_a(vehicle_nr_2)))
}

# Make joint probability table (rows: vehicle X_i, columns: vehicle X_j)
probability_table <- function(vehicle_nr_1, vehicle_nr_2){
  X_i = c(0:max(vehicle_data(vehicle_nr_1)))
  X_j = c(0:max(vehicle_data(vehicle_nr_2)))
  est_P_i = dnbinom(X_i,size=est_r(vehicle_nr_1),mu=est_mu(vehicle_nr_1))
  est_P_j = dnbinom(X_j,size=est_r(vehicle_nr_2),mu=est_mu(vehicle_nr_2))
  prob_table <- array(c(0:max(vehicle_data(vehicle_nr_1))),
                      dim=c(max(vehicle_data(vehicle_nr_1))+1,
                            max(vehicle_data(vehicle_nr_2))+1,1),
                      dimnames = list(c(0:max(vehicle_data(vehicle_nr_1))),
                                      c(0:max(vehicle_data(vehicle_nr_2))),
                      c("Probability Table Xi Xj")))
  for(i in 0:max(vehicle_data(vehicle_nr_1))){
    for(j in 0:max(vehicle_data(vehicle_nr_2))){
      prob_table[i+1,j+1,1] <- joint_distribution(vehicle_nr_1, vehicle_nr_2,i,j)
    }
  }
  return <- prob_table
}

# Sample from Danaher model
sample_Danaherx <- function(vehicle_nr_1, vehicle_nr_2, l){
  prob_table = combinations[[l]]
  grid <- expand.grid(X = 0:max(vehicle_data(vehicle_nr_1)),
                      Y = 0:max(vehicle_data(vehicle_nr_2)))
  grid$p <- c(prob_table)
  for (i in grid$p){
    if (i<0){grid$p[which(grid$p == i)] = 0}
```

```
  }
  smpl <- sample(1:nrow(grid), size = nrow(data), replace=TRUE, prob=grid$p)
  sampled.x <- grid$X[smpl]
}

sample_Danahery <- function(vehicle_nr_1, vehicle_nr_2, l){
  prob_table = combinations[[l]]
  grid <- expand.grid(X = 0:max(vehicle_data(vehicle_nr_1)),
                      Y = 0:max(vehicle_data(vehicle_nr_2)))
  grid$p <- c(prob_table)
  for (i in grid$p){
    if (i<0){grid$p[which(grid$p == i)] = 0}
  }
  smpl <- sample(1:nrow(grid), size = nrow(data), replace=TRUE, prob=grid$p)
  sampled.y <- grid$Y[smpl]
}

# Correlation function
correlation <- function(vehicle_nr_1, vehicle_nr_2){
  w <- est_w(vehicle_nr_1, vehicle_nr_2)
  alpha_1 <- est_a(vehicle_nr_1)
  alpha_2 <- est_a(vehicle_nr_2)
  r_1 <- est_r(vehicle_nr_1)
  r_2 <- est_r(vehicle_nr_2)
  result <- w*(1-exp(-1))^2*((sqrt(r_1*r_2*(1+alpha_1)*(1+alpha_2)))/
                               (alpha_1*alpha_2))*(((alpha_1)/
                                (1+alpha_1-exp(-1)))^(r_1 + 1))*
              (((alpha_2)/(1 + alpha_2 - exp(-1)))^(r_2 + 1))
}
```

A.2. Copula model

A.2.1. Implemented functions in R

```
---
title: "Copula model functions"
author: "Frederiek Backers"
---

```{r, libraries and data import}
library(MASS)
library(fitdistrplus)
library(readxl)
library(feather)
library(plot3D)
library(ggplot2)
library(copula)
library(scatterplot3d)
library(univariateML)
library(VineCopula)
library(sads)
library(discreteRV)
library(plyr)

data <- arrow::read_feather("C:\\Users\\Frederiek.Backers\\Gitlab\\Code
                            \\Rdata_region_Londonzoom.feather")
```
```

```
colnames(data) <- 0:(length(data)-1)
```

```{r, functions}
set.seed(2022)

# Model copula function
model_selectcopula <- function(vehicle_nr_1, vehicle_nr_2){
  data_pobs = pobs(as.matrix(data[c(vehicle_nr_1, vehicle_nr_2)]))
  SelectedCopula <- BiCopSelect(data_pobs[,1], data_pobs[,2], familyset=NA,
                                method="mle")
}
model_Gaussian <- function(vehicle_nr_1, vehicle_nr_2){
  data_pobs = pobs(as.matrix(data[c(vehicle_nr_1, vehicle_nr_2)]))
  Gaussian <- fitCopula(ellipCopula("normal", dim=2), data_pobs, method="ml")
}
model_StudentT <- function(vehicle_nr_1, vehicle_nr_2){
  data_pobs = pobs(as.matrix(data[c(vehicle_nr_1, vehicle_nr_2)]))
  StudentT <- fitCopula(ellipCopula("t", dim=2), data_pobs, method="ml")
}
model_Joe <- function(vehicle_nr_1, vehicle_nr_2){
  data_pobs = pobs(as.matrix(data[c(vehicle_nr_1, vehicle_nr_2)]))
  Joe <- fitCopula(archmCopula("joe", dim=2), data_pobs, method="ml")
}
model_Clayton <- function(vehicle_nr_1, vehicle_nr_2){
  data_pobs = pobs(as.matrix(data[c(vehicle_nr_1, vehicle_nr_2)]))
  Clayton <- fitCopula(archmCopula("clayton", dim=2), data_pobs, method="ml")
}
model_Gumbel <- function(vehicle_nr_1, vehicle_nr_2){
  data_pobs = pobs(as.matrix(data[c(vehicle_nr_1, vehicle_nr_2)]))
  Gumbel <- fitCopula(archmCopula("gumbel", dim=2), data_pobs, method="ml")
}
model_Frank <- function(vehicle_nr_1, vehicle_nr_2){
  data_pobs = pobs(as.matrix(data[c(vehicle_nr_1, vehicle_nr_2)]))
  Frank <- fitCopula(archmCopula("frank", dim=2), data_pobs, method="ml")
}

# Simulate copula function
sim_selcop <- function(vehicle_nr_1, vehicle_nr_2){
  sim <- BiCopSim(N=nrow(data),obj=model_selectcopula(vehicle_nr_1,
                                                      vehicle_nr_2))
}
sim_Gaussian <- function(vehicle_nr_1, vehicle_nr_2){
  sim <- rCopula(n=nrow(data),model_Gaussian(vehicle_nr_1, vehicle_nr_2)@copula)
}
sim_StudentT <- function(vehicle_nr_1, vehicle_nr_2){
  sim <- rCopula(n=nrow(data), model_StudentT(vehicle_nr_1, vehicle_nr_2)@copula)
}
sim_Joe <- function(vehicle_nr_1, vehicle_nr_2){
  sim <- rCopula(n=nrow(data), model_Joe(vehicle_nr_1, vehicle_nr_2)@copula)
}
sim_Clayton <- function(vehicle_nr_1, vehicle_nr_2){
  sim <- rCopula(n=nrow(data),model_Clayton(vehicle_nr_1, vehicle_nr_2)@copula)
}
sim_Gumbel <- function(vehicle_nr_1, vehicle_nr_2){
  sim <- rCopula(n=nrow(data), model_Gumbel(vehicle_nr_1, vehicle_nr_2)@copula)
```

```
}
sim_Frank <- function(vehicle_nr_1, vehicle_nr_2){
  sim <- rCopula(n=nrow(data), model_Frank(vehicle_nr_1, vehicle_nr_2)@copula)
}

# Sample from copula functions
sample_selcopx <- function(vehicle_nr_1, vehicle_nr_2){
  x <- quantile(data[[vehicle_nr_1]], sim_selcop(vehicle_nr_1,vehicle_nr_2)[,1])
}
sample_selcopy <- function(vehicle_nr_1, vehicle_nr_2){
  y <- quantile(data[[vehicle_nr_2]], sim_selcop(vehicle_nr_1,vehicle_nr_2)[,2])
}

sample_Gaussianx <- function(vehicle_nr_1, vehicle_nr_2){
  x <- quantile(data[[vehicle_nr_1]], sim_Gaussian(vehicle_nr_1,
                                                   vehicle_nr_2)[,1])
}
sample_Gaussiany <- function(vehicle_nr_1, vehicle_nr_2){
  y <- quantile(data[[vehicle_nr_2]], sim_Gaussian(vehicle_nr_1,
                                                   vehicle_nr_2)[,2])
}

sample_StudentTx <- function(vehicle_nr_1, vehicle_nr_2){
  x <- quantile(data[[vehicle_nr_1]], sim_StudentT(vehicle_nr_1,
                                                   vehicle_nr_2)[,1])
}
sample_StudentTy <- function(vehicle_nr_1, vehicle_nr_2){
  y <- quantile(data[[vehicle_nr_2]], sim_StudentT(vehicle_nr_1,
                                                   vehicle_nr_2)[,2])
}

sample_Joex <- function(vehicle_nr_1, vehicle_nr_2){
  x <- quantile(data[[vehicle_nr_1]], sim_Joe(vehicle_nr_1, vehicle_nr_2)[,1])
}
sample_Joey <- function(vehicle_nr_1, vehicle_nr_2){
  y <- quantile(data[[vehicle_nr_2]], sim_Joe(vehicle_nr_1, vehicle_nr_2)[,2])
}

sample_Claytonx <- function(vehicle_nr_1, vehicle_nr_2){
  x <- quantile(data[[vehicle_nr_1]], sim_Clayton(vehicle_nr_1,
                                                  vehicle_nr_2)[,1])
}
sample_Claytony <- function(vehicle_nr_1, vehicle_nr_2){
  y <- quantile(data[[vehicle_nr_2]], sim_Clayton(vehicle_nr_1,
                                                  vehicle_nr_2)[,2])
}

sample_Gumbelx <- function(vehicle_nr_1, vehicle_nr_2){
  x <- quantile(data[[vehicle_nr_1]], sim_Gumbel(vehicle_nr_1, vehicle_nr_2)[,1])
}
sample_Gumbely <- function(vehicle_nr_1, vehicle_nr_2){
  y <- quantile(data[[vehicle_nr_2]], sim_Gumbel(vehicle_nr_1, vehicle_nr_2)[,2])
}

sample_Frankx <- function(vehicle_nr_1, vehicle_nr_2){
  x <- quantile(data[[vehicle_nr_1]],sim_Gumbel(vehicle_nr_1, vehicle_nr_2)[,1])
```

```
}
sample_Franky <- function(vehicle_nr_1, vehicle_nr_2){
  y <- quantile(data[[vehicle_nr_2]],sim_Gumbel(vehicle_nr_1, vehicle_nr_2)[,2])
}
```

References

[1] Danaher, P. J. (2007). Modeling Page Views Across Multiple Websites with an Application to Internet Reach and Frequency Prediction. *Marketing Science, 26*(3), 422–437. https://doi.org/10.1287/mksc.1060.0226.

[2] Lee, M. L. T. (1996). Properties and applications of the Sarmanov family of bivariate distributions. *Comm. Statist.: Theory Methods 25*(6) 1207-1222.

[3] Leckenby, J. D., Kishi, S. (1998). Using reach/frequency for Web media planning. *J. Advertising Res. 38*(January) 7-20.

[4] Huang C. Y., Lin. S. (2006). Modeling the audience banner ad exposure for internet advertising planning. *J. Advertising 35*(2) 23-37.

[5] Wood, L. (1998). Internet ad buys - What reach and frequency do they deliver? *J. Advertising Res. 38*(January) 21-28.

[6] Delignette-Muller, M. L. Dutang, C. (2015). fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software, 64*(4). https://www.jstatsoft.org/.

[7] Klopper, J. H. (2018, 5 december). *Exact test of goodness of fit.* Rpubs. Accessed at 20 June 2022, from https://rpubs.com/juanhklopper/exact_test_of_goodness_of_fit.

[8] Hofert, M., Maechler, M., Yan, J., Neslehovà, J.G. Morger, R. (2022). *Package "copula": Multivariate Dependence with Copulas* (1.1-0) [R software package]. https://cran.r-project.org/web/packages/copula/copula.pdf.

[9] Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B. Erhardt, T. (2022). *Package "VineCopula": Statistical Inference of Vine Copulas* (2.4.4) [R software package]. https://cran.r-project.org/web/packages/VineCopula/VineCopula.pdf.

[10] Bakker, B. J. (2020, juli). *Modelling finite mixture joint distributions.* Delft University of Technology. https://repository.tudelft.nl.

[11] Schmidt, T. (2006). *Coping with Copulas.* University of Leipzig. https://web.archive.org/web/20100705040514/http://www.tu-chemnitz.de/mathematik/fima/publikationen/TSchmidt_Copulas.pdf.

[12] Bolbolian Ghalibaf, M. (2020). Relationship Between Kendall's tau Correlation and Mutual Information. *Revista Colombiana de Estadística, 43*(1), 3-20. https://doi.org/10.15446/rce.v43nl.78054.

[13] Chang, B. (2019, 8 mei). *Copula: A Very Short Introduction.* Bochang.me. Accessed at 20 June 2022, from https://bochang.me/blog/posts/copula/.

[14] Nelsen, R. B. (2006). *An Introduction to Copulas* (2nd edition). Springer.

[15] Ehrenberg, A. S. C. (1992). *Repeat Buying: Facts, Theory and Applications* (2nd edition). Charles Griffin and Company Limited, London.

[16] Morrison, D. G., Schmittlein D. C. (1988). Generalizing the NBD model for costomer purchases: What are the implications and is it worth the effort? *J. Bus. Econom. Statist. 6*(2, april) 145-159.

[17] Zhou, Y. Wu, C. (2018). Multivariate frequency analysis of urban rainfall characteristics using three-dimensional copulas. *Water Science Technology.* https://doi.org/10.2166/wst.2018.103.

[18] Ma, D. (2011, 29 juli). *The Negative Binomial Distribution.* A Blog on Probability and Statistics. Accessed at 25 April 2022, from https://probabilityandstats.wordpress.com/tag/poisson-gamma-mixture/.

[19] H. (2015, 5 februari). *Goodness of fit test in R.* R-Bloggers. Accessed at 20 June 2022, from https://www.r-bloggers.com/2015/01/goodness-of-fit-test-in-r/.

[20] Danaher, P. J. Smith, M. S. (2011). Modeling Multivariate Distributions Using Copulas: Applications in Marketing. *Marketing Science, 30*(1), 4-21. https://doi.org/10.1287/mksc.1090.0491.

[21] Rijksuniversiteit Groningen Albers, C. (z.d.). *Estimating bivariate distributions assuming some form of dependence.* https://fse.studenttheses.ub.rug.nl/8655/1/Math_Drs_1998_CAlbers.CV.pdf.

[22] *Kendall's Tau.* (2017, 14 november). Statistics How To. Accessed at 2 June 2022, from https://www.statisticshowto.com/kendalls-tau/.

[23] M. (2016, 29 maart). *How to fit a copula model in R [heavily revised]. Part 2: fitting the copula.* R-Bloggers. accessed at 11 of June 2022, from https://www.r-bloggers.com/2016/03/how-to-fit-a-copula-model-in-r-heavily-revised-part-2-fitting-the-copula/.

[24] Valle, L. D., Leisen, F. Rossini, L. (2017). Bayesian non-parametric conditional copula estimation of twin data. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 67*(3), 532-548. https://doi.org/10.1111/rssc.12237.