

Department of Statistics, University of Auckland

Summer Scholarship 2015-2016

Analysing social network data from the New Zealand General Social Survey 2014

Name: David Ka En Chan

Supervisor: Roy Lay-Yee (COMPASS), Brian McArdle (Statistics)

Degree: Bachelor of Music and Science

Host Unit: COMPASS Research Centre

Thanks to Professor Peter Davis and the COMPASS Research Centre for hosting me, and to Statistics New Zealand for access to the New Zealand General Social Survey data.

Disclaimer

Access to the data presented was managed by Statistics New Zealand under strict micro-data access protocols and in accordance with the security and confidentiality provisions of the Statistic Act 1975. Our findings are not Official Statistics. The opinions, findings, recommendations, and conclusions expressed are those of the researcher, not Statistics NZ.

Summary

The New Zealand General Social Survey (NZGSS) is a cross-sectional survey undertaken every two years by Statistics NZ. The two primary aims of my project were: to derive measures describing social network data in the NZGSS 2014; and to use these measures to assess the effect of social networks on well-being outcomes. There were survey design issues which I needed to tackle before exploratory analysis could be carried out. Given the nature of the project aims, I choose to use multivariate analysis to analyse the dataset. I used the concept of model selection to determine the final model of the multivariate analysis. In conclusion, the approach of assessing the effect of social networks on well-being outcomes needs continued refinement of what has been explored here.

Abstract

The New Zealand General Social Survey (NZGSS) is a cross-sectional survey undertaken every two years by Statistics NZ. The two primary aims of my project were: to derive measures describing social network data in the NZGSS 2014; and to use these measures to assess the effect of social networks on well-being outcomes. I employed multivariate analysis to hypothesise whether the social network measures derived from survey questions in the NZGSS 2014 can predict a selection of well-being measures. Also, a design matrix is required for the statistical analysis because of how the NZGSS 2014 survey questionnaire was mapped out structurally. The design matrix is made up of conditioning variables, which partition the sample into subgroups. From the five conditioning variables that were identified, we kept three of them in the analysis. Using redundancy analysis, a multivariate analysis technique, we were able to identify a selection of social network measures and demographic measures to begin a process known as model selection. Model selection is a process of validating whether a variable contributes to a statistical model, in this case the redundancy analysis output, by evaluating the goodness of fit. The final model contained three conditioning variables and 47 other variables, where some of the variables were partitioned by the conditioning variables. This project has explored one possibility to assess the effect of social networks on well-being outcomes. In conclusion, the approach outlined in this report needs continued refinement.

Table of Contents

Summary	ii
Abstract.....	iii
List of figures and tables	v
1.0 Introduction	1
1.1 Background	1
1.2 Research Aims.....	2
2.0 Methods.....	3
2.1 Data manipulation - the design matrix	3
2.2 Data manipulation - other	4
2.3 Redundancy analysis – a multivariate analysis method	4
2.4 Model selection	5
3.0 Results.....	6
3.1 Exploratory analysis	6
3.2 The full model	10
3.3 The baseline model.....	10
3.4 The final model	11
4.0 Discussion.....	13
References	15
Appendices.....	16
Appendix A.....	16
Appendix B	16
Appendix C	16
Appendix D.....	16
Appendix E	18
Appendix F	18

List of figures and tables

Figure 1	1
Figure 2	3
Figure 3	6
Figure 4	7
Figure 5a.....	7
Figure 5b	8
Figure 5c.....	8
Figure 6	9
Figure 7	11
Figure 8	12
Table 1.....	9
Table 2.....	10
Table 3.....	10
Table 4.....	11

1.0 Introduction

1.1 Background

The New Zealand General Social Survey (NZGSS) is a cross-sectional survey undertaken every two years by Statistics New Zealand. The NZGSS series began in 2008 and consists of 13 modules which focus on various domains of well-being. The population of interest for the NZGSS are New Zealand residents currently living in a private dwelling aged 15 or older. With each NZGSS, a supplementary module is included looking at a particular aspect of New Zealanders' lives.

The NZGSS 2014 had an 80.3% response rate, which yielded a sample size of approximately 8800 respondents. Sampling methodology and reasoning can be found on the Statistics New Zealand webpage for the NZGSS 2014 (MacPherson, 2015). My project in particular focuses on the NZGSS 2014's supplementary module on social networks and support. The module itself is split into three separate sections: characteristics of social networks; strength of social networks; and effectiveness of social networks.

Social network data are commonly analysed by the social sciences. The population of interest are known as actors and by definition can be a part of any other actor's network. The social sciences use a paradigm of techniques coined as social network analysis (Scott & Carrington, 2011). From the academic literature I have read, there was nothing which covered the possibility of using social network data to predict other variable(s). There was however information on how to gather social network data through surveys (Marsden, 2011). As the NZGSS 2014 uses a representative sample of the population, it is appropriate to collect an actor's social network data as an egocentric network (Statistics New Zealand, 2015a). We define an 'ego' to be the observed actor who is placed in the middle of their network. We also define an 'alter' to be a participant of an ego's network.

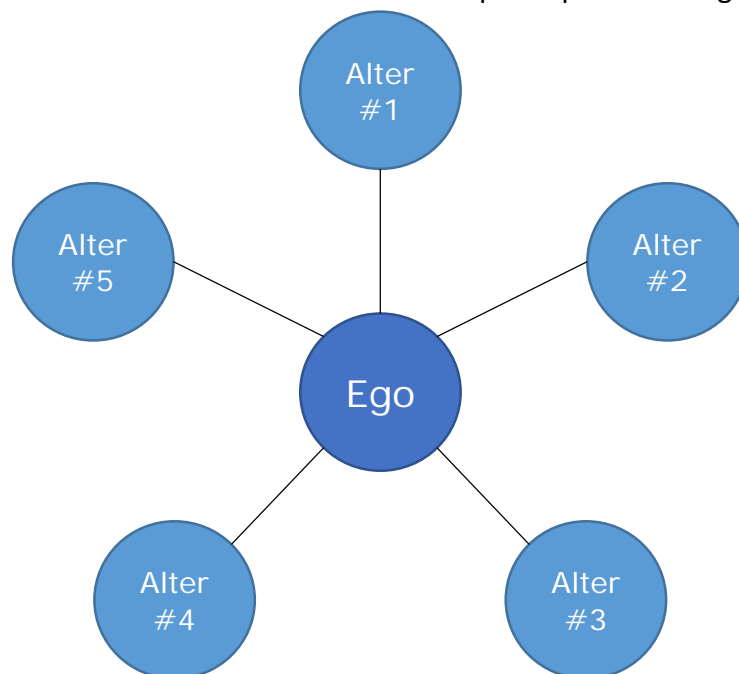


Figure 1: An egocentric network

1.2 Research Aims

The two primary aims of my project were: to derive measures describing social network data in the New Zealand General Social Survey 2014 (NZGSS 2014); and to use these measures to assess the effect of social networks on well-being outcomes.

The main obstacle with NZGSS 2014 social network data is the lack of consistency for alter information. This makes it difficult to derive social network measures for egocentric networks. Not only that, most established social network analysis techniques purely analyse the social network structure and improving it with well-being factors. Whereas I want to use social network measures to assess their effect on well-being outcomes. This is why I am going to use multivariate analysis to analyse the social network measures in the NZGSS 2014.

I employed the use of multivariate analysis to hypothesise whether the social network measures derived from survey questions in the NZGSS 2014 can predict a selection of well-being outcomes. Multivariate analysis aims to reduce the dimensionality of a high dimension dataset. It generally does this by taking advantage of the correlation between the response variables. We can then use a series of multivariate techniques to explore the reduced dimensionality numerically or graphically. For the project I was advised to limit the response variables to two well-being measures: self-reported overall life satisfaction, and self-reported sense of purpose. These were the two primary well-being measures reported by Statistics NZ (MacPherson, 2015). This also enables the possibility to add more well-being measures for further research. The methodology section below will outline how I approached multivariate analysis of the social network data.

had to be partitioned into subgroups, which inflated the size of the X matrix. If a conditioning variable was to be removed, the associated respondents would also be removed from the dataset.

2.2 Data manipulation - other

There were three other data manipulation phases to be conducted with the NZGSS 2014 dataset. These data manipulation phases were carried out continually throughout the project.

The number of respondents to be used from the NZGSS 2014 dataset was determined by two definitions. As outlined in Section 2.1, the sample size for analysis was controlled by the number of conditioning variables kept. The sample size was also controlled by whether a respondent was able to answer the question “qSupportType”, without selecting either “Don’t Know” or “Refused” as their response (variable description can be found in Appendix A).

The second data manipulation phase was to clean the dataset. This meant ensuring that each variable was coded correctly either as a factor variable or a numeric variable. I also needed to check whether each factor variable had an appropriate baseline level, and if not, implement it into the variable. Lastly, I needed to appraise whether the responses of certain questions could be collapsed into another similar question. An example of this is the following set of questions: “qOneFamTalk”; “qOneFamVideo”; “qOneFamPhone”; “qOneFamWrite”; “qFamTalk”; “qFamVideo”; “qFamPhone”; and “qFamWrite”. I can collapse the responses of “qOneFamTalk”, “qOneFamVideo”, “qOneFamPhone”, and “qOneFamWrite” into “qFamTalk”, “qFamVideo”, “qFamPhone”, and “qFamWrite” respectively (variable descriptions can be found in Appendix A).

The third data manipulation phase related to whether there was a need to derive further social network measures from the collected data. Creating these derived social network measures may involve the addition and/or removal of variables, as it might introduce co-linearity. The inclusion of the derived network measures will be informed by Marsden (2011).

2.3 Redundancy analysis – a multivariate analysis method

It must be established that redundancy analysis is an extension of multivariate regression; that it is a method where we can make sense of the fit of the regression model. Multivariate regression is simply like univariate regression. In the univariate space we regress a Y variable against an X matrix. In the multivariate space we instead regress a Y matrix against an X matrix. Mathematically a multivariate regression model is essentially a collection of univariate regression models produced for each Y variable in the Y matrix. We could try to interpret the output as is, but if the dataset is of high dimensionality it becomes rather difficult. Hence the multivariate nature of the regression model is associated with assisting the interpretation of the output.

As with other established multivariate techniques, redundancy analysis reduces the dimensionality of the fitted Y matrix generated by the multivariate regression model (Van den Wollenberg, 1977). The methodology behind how redundancy analysis reduces the dimensionality is similar to a principal components analysis. The interpretation of redundancy analysis axes is quite similar to the interpretation of principal component axes. Redundancy analysis scores are also generated by regressing the fitted Y matrix against the X matrix. My personal interpretation of redundancy

analysis is that it could be likened to analysing the residual structure of a regression model. I used a dedicated R package called `vegan` to produce the redundancy analysis objects for the project. The model selection process (Section 2.4) will be powered by models built from the redundancy analysis objects in R.

2.4 Model selection

Given the sheer number of social network measures (questions) in the NZGSS 2014 dataset, if we were to start the variable elimination process from the full model, it would be time inefficient. So I executed analysis of variance on the marginal effects of variables when compared against the residual of the full model (Oksanen, 2016). Since the main goal of this analysis of variance was to define a baseline model to construct our multivariate regression model, I used 999 permutations per variable (the default setting) to generate the associated summary statistics. The chosen number of permutations meant that only variables with p-values less than 0.05 were considered for the baseline model. The upper limit for this consideration is driven by a rule-of-thumb commonly used in statistics. It should also be noted that we included a handful of socio-demographic measures (to go alongside the social network measures) including: age; gender; ethnicity; labour force status; Maori descent; highest qualification; born in New Zealand; number of years in New Zealand for migrants; presence of a benefit income source; and presence of a non-benefit income source.

With the base model identified by the analysis of variance, the next step was to add variables or remove existing variables which I suspected should increase the goodness of fit of the multivariate regression model. If an increase was not achieved, an informed choice was made based on the changes in the goodness of fit on a variable's inclusion or removal. The main goodness of fit statistic that I used in my model selection process was adjusted R^2 . Where R^2 describes the variance captured by a redundancy analysis axis, adjusted R^2 has the same functionality, except that it also applies a penalty which is based on the number of variables in the X matrix

This phase of the model selection was largely driven by methods described by Marsden (2011), and documentation that Statistics NZ has released regarding their own analysis of the social network data collected by the NZGSS 2014 (Statistics New Zealand, 2015b). It is important to note that there is not an ideal goodness of fit threshold to be satisfied.

3.0 Results

3.1 Exploratory analysis

The first phase of trimming respondents from the dataset dropped our sample size to 6189 respondents from approximately 8800 original respondents (outlined in Section 2.2). For our response variables - self-reported overall life satisfaction and self-reported sense of purpose - the following conditioned univariate plots show an interesting pattern.

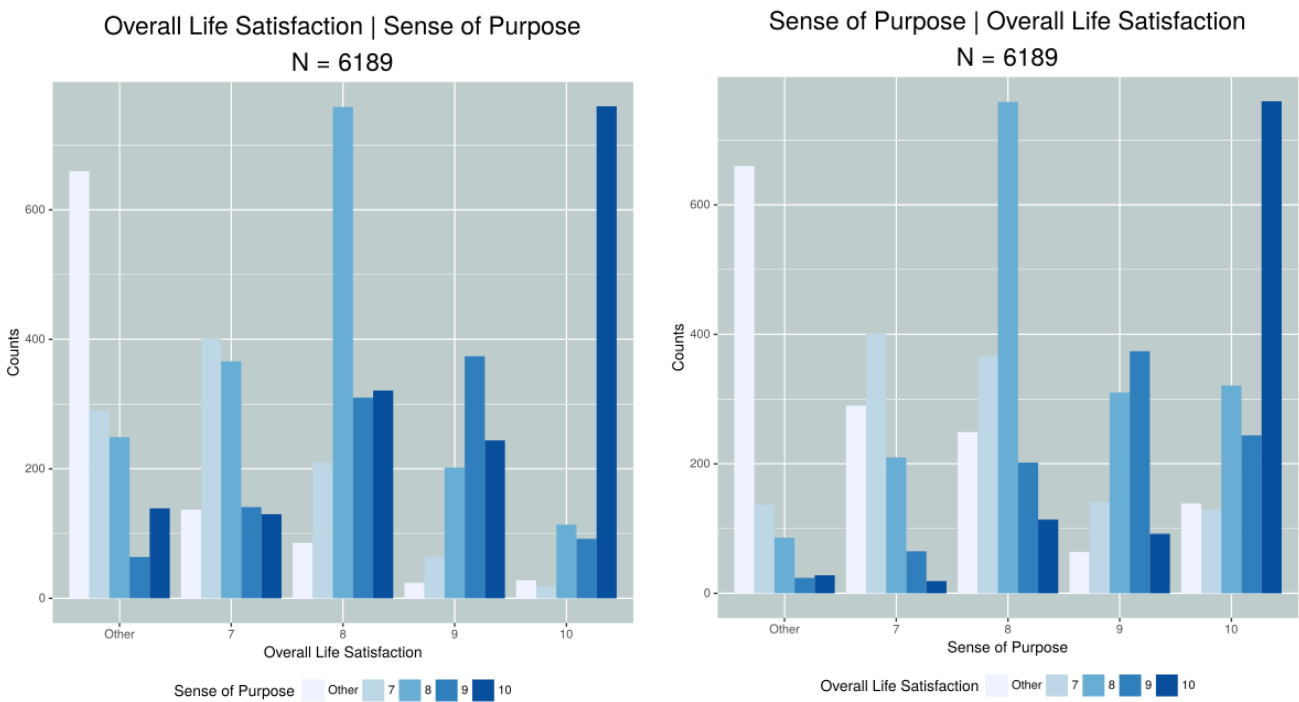


Figure 3: Univariate displays of self-reported Overall Life Satisfaction and self-reported Sense of Purpose

I opted to use a reduced scale for both of the response variables in order to reduce visual clutter in the plots. The univariate plots suggests that there is a correlation structure between the two self-reported well-being measures. In order to visually explore the correlation structure further, I have chosen to use a bubble plot. A bubble plot is an extension of a scatterplot with an additional continuous variable. This additional continuous variables controls the sizes of the bubbles. I have opted to use the univariate plots' reduced scale again for the bubble plot.

For the following bubble plot (Figure 4), the third dimension (size) represents the density of respondents for each possible reduced scale combination. To further assist the comparison between bubble sizes, the density is also represented by a scale of heat-map colours ranging from yellow to red.

The bubble plot (Figure 4) confirms our hypothesis of a correlation structure between the two self-reported well-being measures. Hence, a redundancy analysis will be sufficient to meet our two project aims: to derive measures describing social network data in the NZGSS 2014; and to use these measures to assess the effect of social networks on well-being outcomes.

Overall Life Satisfaction by Sense of Purpose

$R^2 = 0.3442$, $N = 6189$

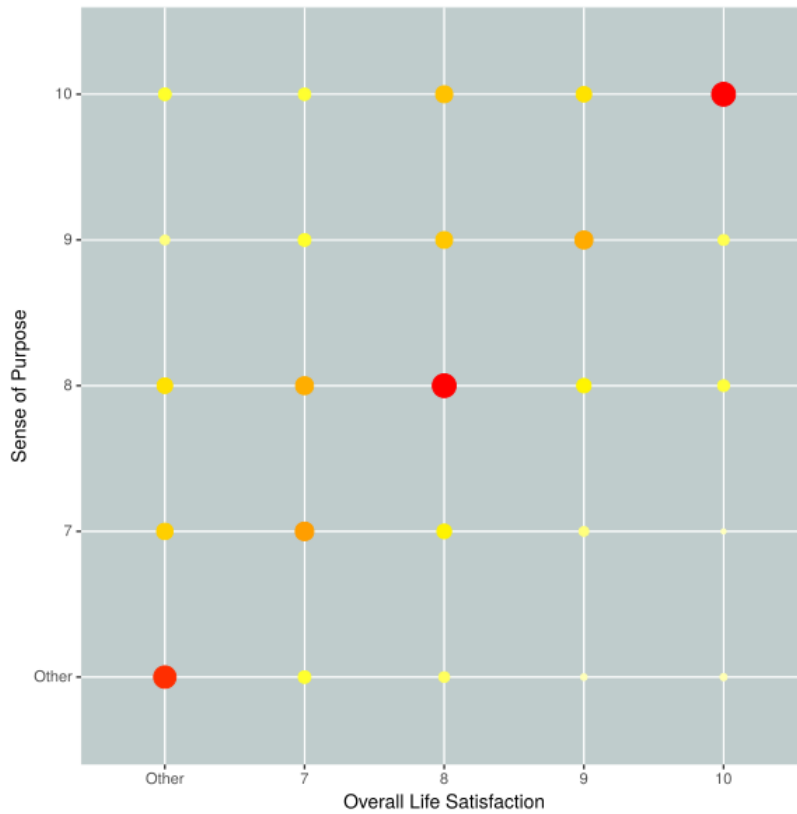


Figure 4: A bubble plot of self-reported Overall Life Satisfaction against self-reported Sense of Purpose

The effect of the conditioning variables on the NZGSS 2014 dataset was quite interesting. I had to make a number of judgement calls on which of the conditioning variables should be kept. My two supervisors assisted me on this task.



Figure 5a: A visual exploration of the conditioning variables

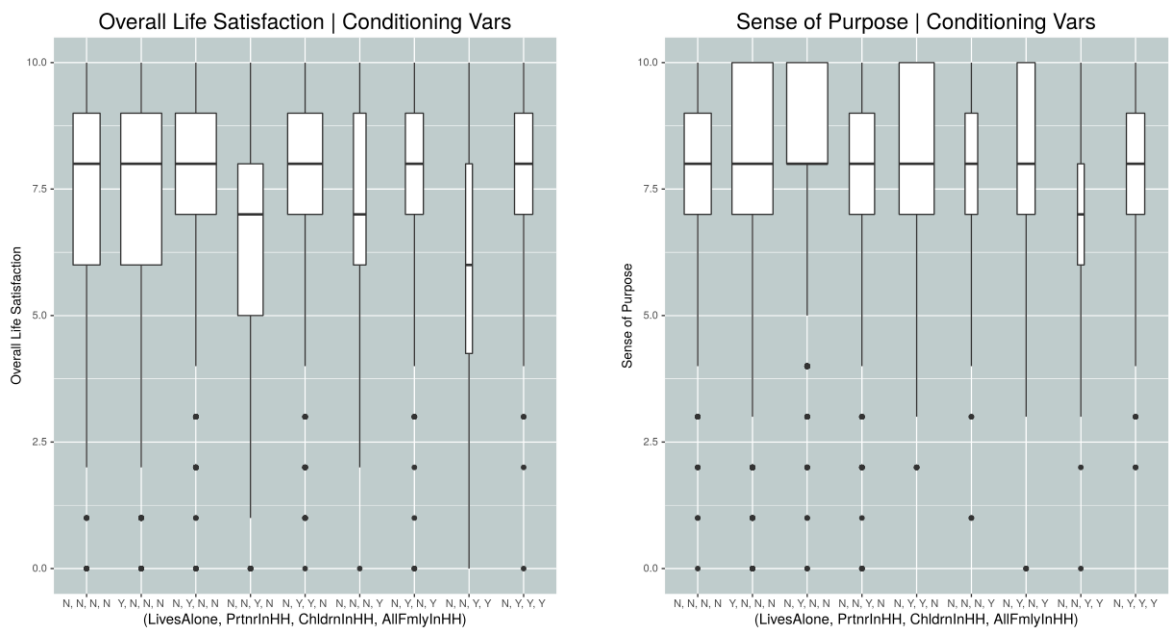


Figure 5b: Continued visual exploration of conditioning with AllFrndInHH dropped

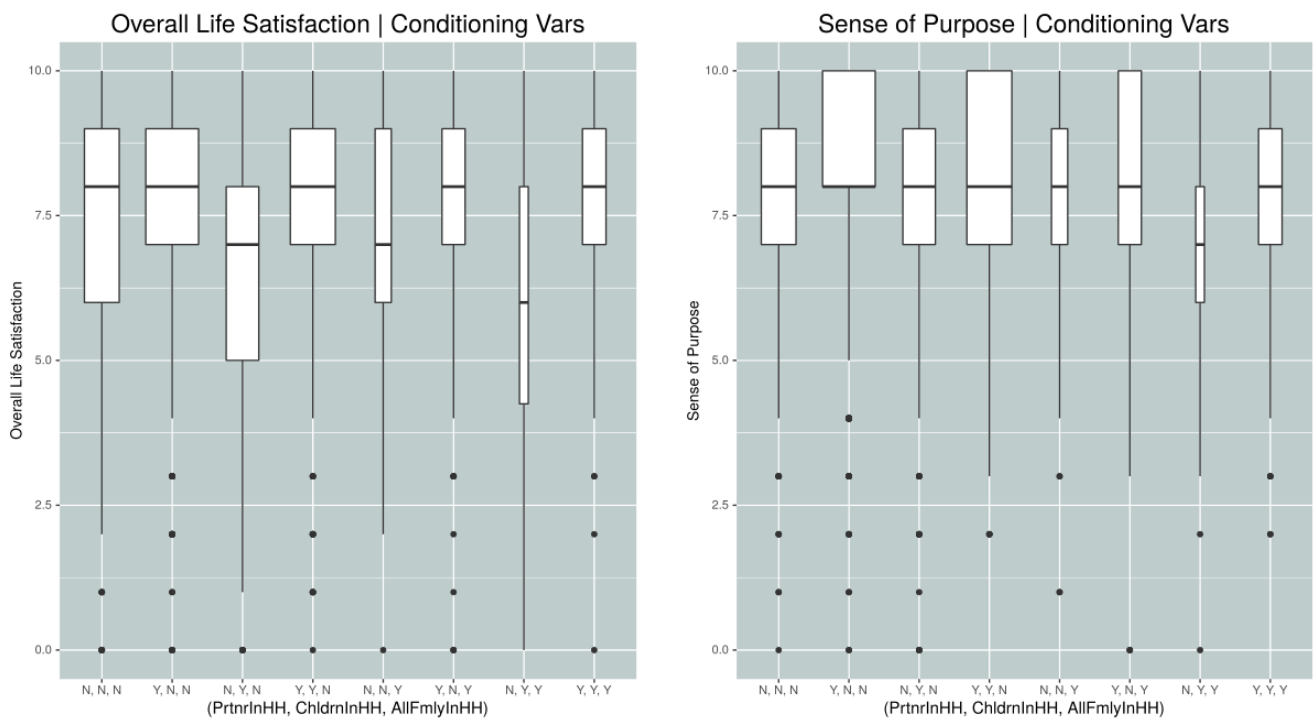


Figure 5c: Final visual exploration of the conditioning with both LivesAlone and AllFrndInHH dropped

The preceding boxplots visually explore how the conditioning variables partition the sample. The width of the individual boxplots are determined by the number of respondents in the corresponding subgroup. Figure 5a shows that the subgroups which met the condition for “AllFrndInHH” only made up a small partition of the sample. Hence I dropped “AllFrndInHH” as a conditioning variable, along with its associated subgroups. Figure 5b reveals that the four remaining conditioning variables seem fine. However, it suggests that it is possible to further simplify our redundancy analysis by dropping “LivesAlone” as a conditioning variable. This is because, by

definition, meeting the condition for “LivesAlone” locks out the possibility to meet the requirements for the other conditioning variables. I performed a t.test of differences in order to determine whether I should drop “LivesAlone”. (Conditioning variable descriptions can be found in Appendix A).

	t-statistic	df	p-value
Overall Life Satisfaction	0.4876	2108	0.6259
Sense of Purpose	0.8029	1588	0.4222

Table 1: T.test on the differences between the subgroup with the condition “LivesAlone” satisfied and the subgroup with all conditioning variables unsatisfied

There were no differences in either of the response variables for the two tested subgroups (p-values: 0.6259 and 0.4222 respectively). Hence I dropped “LivesAlone” as a conditioning variable with its associated subgroup. Figure 5c shows the visual partitioning of the three remaining conditioning variables, which brings our final sample size to 4641. The analysis of variance output can be found in Appendix B, which reveals that the remaining conditioning variables were highly significant.

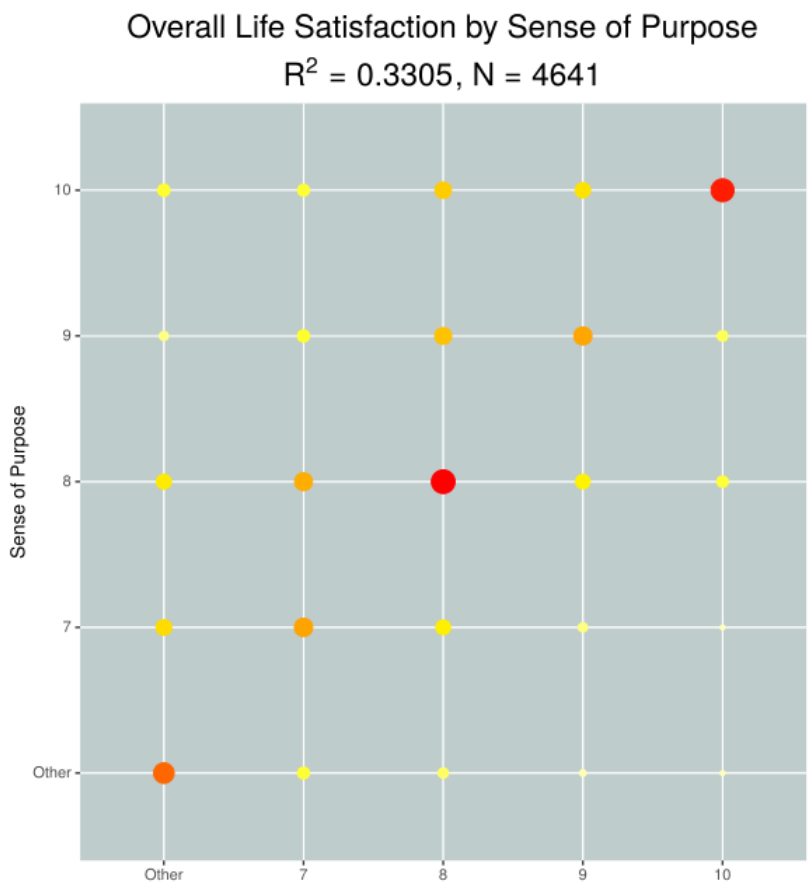


Figure 6: A bubble plot of self-reported Overall Life Satisfaction against self-reported Sense of Purpose on the reduced dataset

The final bubble plot reveals that the correlation structure still existed with the reduced sample size (N = 4641). With the conclusion of the exploratory analysis, I proceed to the redundancy analysis.

3.2 The full model

The redundancy analysis on the full model yielded the following summary statistics. This model contained all of the social network measures and assorted socio-demographic variables.

	RDA1	RDA2	Overall Life Satisfaction (OLS)	Sense of Purpose (SoP)
RDA1	1.0000	-0.1610	0.8252	0.8649
RDA2	-0.1610	1.0000	0.4247	-0.6346
OLS	0.8252	0.4247	1.0000	0.4302
SoP	0.8649	-0.6346	0.4302	1.0000
Pseudo-F: 1.983 (with 684, 3954 Degrees of Freedom)				
Significance: 0.029				
R ² : 0.2554		Adjusted R ² : 0.1266		

Table 2: Summary statistics of the full model

Pseudo-F was quite low, which can be explained by the extremely large number of variables in the X matrix. The variation captured by the redundancy analysis axes was high ($R^2 = 0.2554$). However, the penalty in variation captured for having a large X matrix, adjusted R^2 , was quite low in comparison with the R^2 which was not satisfactory (adjusted $R^2 = 0.1266$). Surprisingly, the redundancy analysis axes (RDA1 and RDA2) correlate reasonably well with our two response variables.

3.3 The baseline model

The redundancy analysis on the baseline model yielded the following summary statistics. The model contained the significant variables detected by the marginal effects analysis of variance within the X matrix of the full model (p -value < 0.05). Some of the least significant variables were included because they were related to the significant variables. The analytical output for this analysis of variance can be found in Appendix C.

	RDA1	RDA2	Overall Life Satisfaction (OLS)	Sense of Purpose (SoP)
RDA1	1.0000	0.2320	0.8535	0.8375
RDA2	0.2320	1.0000	-0.3089	0.7258
OLS	0.8535	-0.3089	1.0000	0.4302
SoP	0.8375	0.7258	0.4302	1.0000
Pseudo-F: 5.466 (with 171, 4467 Degrees of Freedom)				
Significance: 0.006				
R ² : 0.1730		Adjusted R ² : 0.1414		

Table 3: Summary statistics of the baseline model

Pseudo-F was significantly higher than in the full model, which can be explained by the reduction in variables in the X matrix. There was improvement in the adjusted R^2 for the RDAs (adjusted $R^2 = 0.1414$), even though the model's R^2 decreased ($R^2 = 0.1730$). It is important to note that for the model selection process, we are ultimately interested in improvements of adjusted R^2 . Also, it is noticeable that RDA1 now correlated better with self-reported Overall Life Satisfaction rather than self-reported Sense of Purpose.

3.4 The final model

The redundancy analysis on the final model yielded the following summary statistics. This model contained most of the variables found in the baseline model, and additional variables which were tested with model selection. A full list of the selected variables can be found in Appendix D.

	RDA1	RDA2	Overall Life Satisfaction (OLS)	Sense of Purpose (SoP)
RDA1	1.0000	0.2244	0.8504	0.8408
RDA2	0.2244	1.0000	-0.3219	0.7163
OLS	0.8504	-0.3219	1.0000	0.4302
SoP	0.8408	0.7163	0.4302	1.0000
Pseudo-F: 4.902 (with 203, 4435 Degrees of Freedom)				
Significance: 0.004				
R ² : 0.1883			Adjusted R ² : 0.1459	

Table 4: Summary statistics of the final model

Pseudo-F did not drop significantly with conclusion of the model selection process. The adjusted R² only increased marginally (adjusted R² = 0.1459). However, the R² for the final model was reasonable without compromising the adjusted R² (R² = 0.1883). There was an absence of a change in the correlation structure between the RDAs and the two response variables.

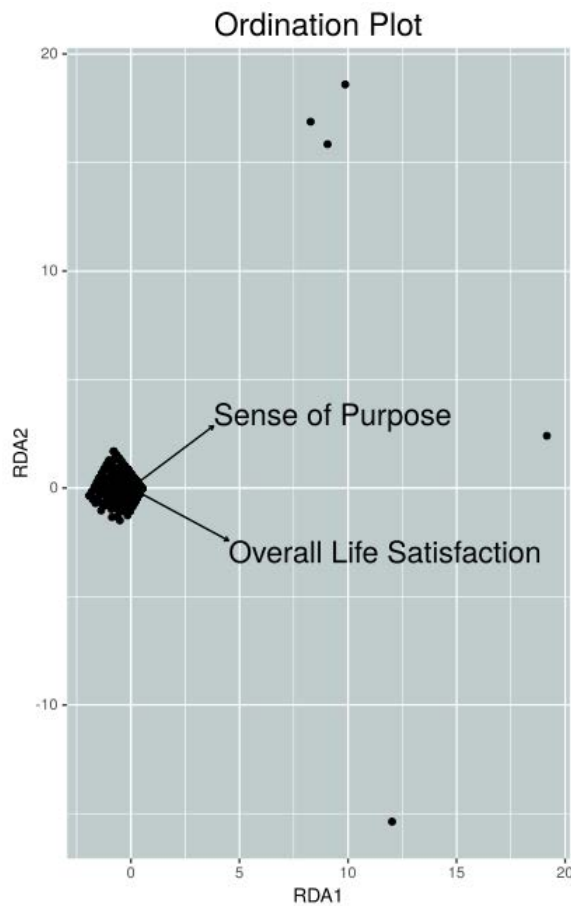


Figure 7: Ordination plot of the redundancy analysis model

The ordination plot of the final model seemed unusual due to the fact that our Y matrix consisted of only two response variables. The outliers present in Figure 7 can be attributed to respondents who did not rate at least one well-being measure. This is made even more apparent once we generated bubble plots of the response variables as the colour and size for each respondent.

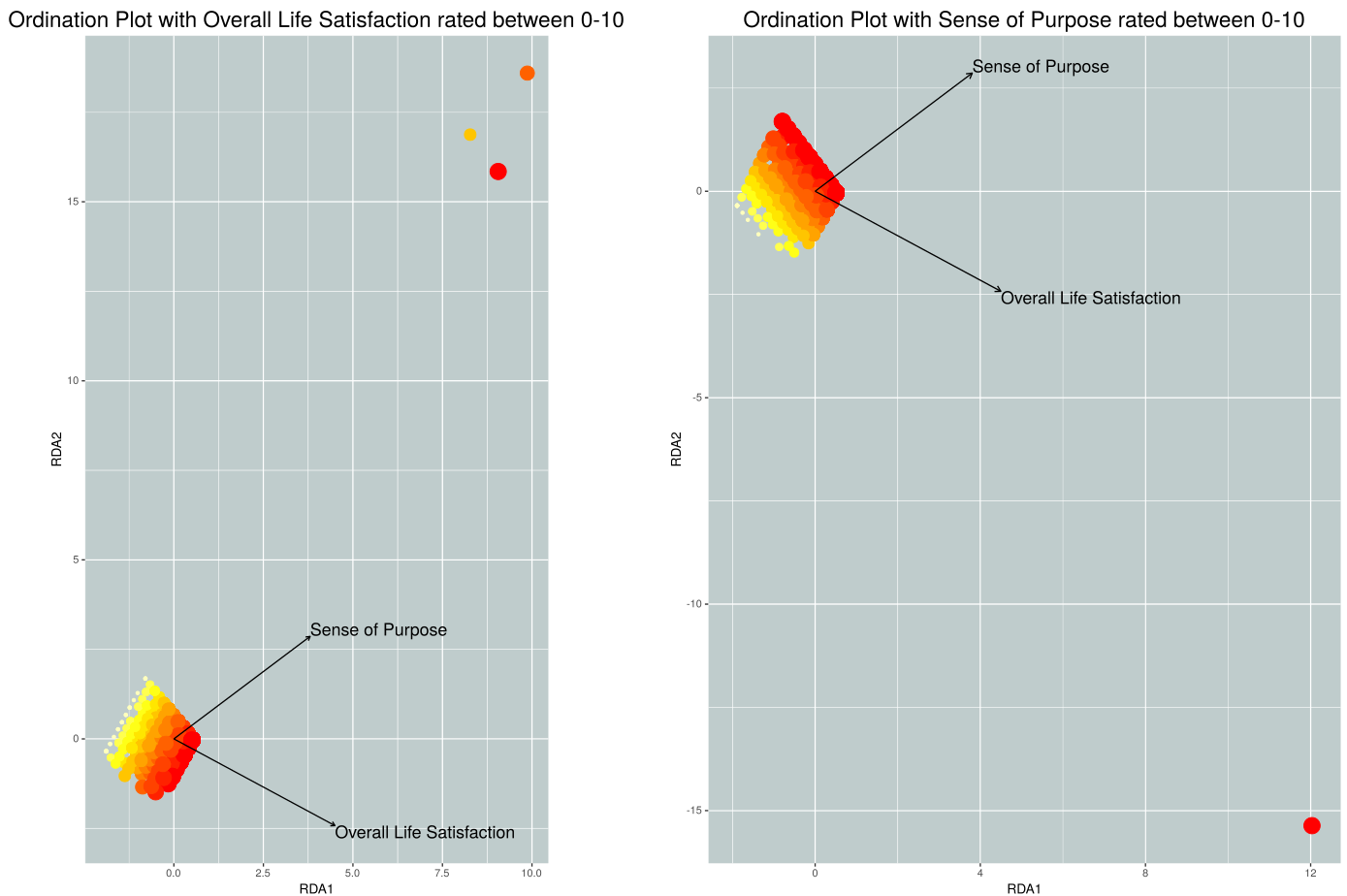


Figure 8: Bubble plots of the redundancy analysis model for each variable in the Y matrix

As already suggested by the two biplot arrows produced for the response variables in Figure 7, Figure 8 reveals a clear planar effect for both of the response variables in the RDA space. It would be interesting to explore how additional response variables in the Y matrix would affect the data-cloud generated by inspecting the two primary RDAs of the redundancy analysis.

It was possible to use a rank-order check on the biplot arrows generated for the X matrix, if they were reported as Euclidean distances (Appendix E). Variables with biplot arrows of Euclidean distances greater than 0.2, were largely associated with the partitioned continuous social network measures. These identified measures come from a family of social network questions which allow the respondent to evaluate their social network in relation to the most impactful life changing event. Other social network measures listed include several 'position generator' questions (Marsden, 2011) which were coded as factors in R. Socio-demographic measures listed included the respondent's age, whether the respondent had a benefit based income, and whether the respondent had a non-benefit based income.

4.0 Discussion

As stated in Section 1.2, the two primary aims of my project were: to derive measures describing social network data in the NZGSS 2014; and to use these measures to assess the effect of social networks on well-being outcomes.

In the final model (Section 3.4) we derived five new measures describing the collected social network data: “SNCqFamRelations_Close”; “SNCqFrndHomophily_Var”; “SNCqFrndHomophily_Max”; “SNCqClubMembership”; and “EFFqAlterMatchScore” (descriptions can be found in Appendix D). Each of these five measures improved either the layman interpretability of the X matrix or improved the statistical fit of the final model. Two of the five derived measures- “SNCqFrndHomophily_Var” and by definition “SNCqFrndHomophily_Max” (Marsden, 2011) - describe a series of ‘position generator’ questions found in the NZGSS 2014 regarding the homophily of one’s friends. I coded “SNCqFrndHomophily_Var” as a factor variable in R because it was more interesting to look at whether respondents were similar to their friend social network or not. In order to increase the interpretability of “SNCqFrndHomophily_Var”, “SNCqFrndHomophily_Max” was also included. “EFFqAlterMatchScore” for a significant portion of respondents had a score of zero, and a subset of that portion were defined as such because of the survey design for the social network module. This is because the derivation relied on respondents being able to answer the “qFirstSupport” questions, as respondents could finish the survey by the nature of its design before that point (see Appendix A for variable description). This poses a question on whether “EFFqAlterMatchScore” should be kept for further research. These five measures complement the first research aim, together with the other social network measures found in Appendix D.

To summarise the entirety of Section 3.0, we were able to construct a redundancy analysis model to assess the effect of social networks on well-being outcomes. A swathe of social network measures and a handful of socio-demographic measures were deemed to be statistically significant. These identified measures found in Appendix D can serve as a foundation for future analyses. Further improvements in the identified measures can be undertaken alongside the continued development of egocentric social network analysis techniques. However, these plausible improvements will be constrained by NZGSS 2014’s survey design.

Extending the point made about outliers in Section 3.4, their removal was investigated to improve the goodness of fit. However, it produced similar summary statistics and ordination plots of the RDAs. One should note that because the outliers were removed after model selection, there was a possibility that different social network measures could have been included in the final model. Further research into this would look into identifying potential outliers and removing them, before conducting redundancy analysis and the model selection process.

The inclusion of other well-being measures in the Y matrix is an opportunity to further tune the final model produced by the redundancy analysis. The main goodness of fit statistic we would be interested would be the correlations between the RDAs and the well-being measures in the Y matrix. This is because comparisons of R^2 or adjusted R^2 are redundant, due to the inclusion of extra well-being measures. This further work could be carried out in R using the `vegan` package,

adapting the work carried out from this project (Appendix F). At this time, it is unknown whether there is a similar implementation of `vegan` in SAS.

Finally I want to discuss the redundancy analysis technique and the conditioning variables. Redundancy analysis did indeed meet the aims of the project well. The main concern with redundancy analysis is that social network data is not inherently mathematical. Hence the marginal analysis of variance technique, used in Section 3.3 to identify the baseline model, will always detect the variables which contribute the most mathematically with the redundancy analysis technique. This makes the model selection process quite laborious, because adjusted R^2 and the Pseudo-F statistic are the most reliable summary statistics we can use (see Appendix F). Redundancy analysis seems to be the most appropriate statistical analysis technique for these data. The implementation of conditioning variables in the NZGSS 2014 dataset was fine. This did open up some insight on which continuous variables were significant to which partitions of the sample. It is still intuitively interpretable as well in the analytical output. Further investigation should be done to determine the right selection of conditioning variables, because it will affect the sample size. This could be done with the assistance of Statistics New Zealand, because they have designed the NZGSS 2014. Also, one could alleviate the power of conditioning variables by collapsing, deriving, or recoding the social network data collected in the NZGSS 2014.

This project has explored one possibility to assess the effect of social networks on well-being outcomes. Applying the same methods to other social network datasets which contain well-being measures may not be appropriate. This is primarily driven by the fact that I have used redundancy analysis to determine which social network measures were statistically significant. Also, the NZGSS 2014's survey design made it infeasible to use established social network analysis techniques. In conclusion, this approach of assessing the effect of social networks on well-being outcomes needs continued refinement of what I have outlined in this report.

References

- MacPherson, L. (2015). *New Zealand general social survey 2014: Key facts*. Retrieved 02/04, 2016, from http://www.stats.govt.nz/browse_for_stats/people_and_communities/Households/nzgss_HO_TP2014.aspx
- Marsden, P. V. (2011). Survey methods for network data. In J. Scott, & P. J. Carrington (Eds.), *The SAGE handbook of social network analysis* (pp. 370-388). London; Thousand Oaks, Calif.: SAGE.
- Oksanen, J. (2016). *Package 'vegan'*. Retrieved 02/09, 2016, from <https://cran.r-project.org/web/packages/vegan/vegan.pdf>
- Scott, J., & Carrington, P. J. (2011). Introduction. In J. Scott, & P. J. Carrington (Eds.), *The SAGE handbook of social network analysis* (pp. 1-8). London; Thousand Oaks, Calif.: SAGE.
- Statistics New Zealand. (2015a). *New zealand general social survey 2014: Data quality*. Retrieved 02/04, 2016, from http://www.stats.govt.nz/browse_for_stats/people_and_communities/Households/nzgss_HO_TP2014/Data%20Quality.aspx#Accuracy2
- Statistics New Zealand. (2015b). *Social networks and support*. Retrieved 02/09, 2016, from http://www.stats.govt.nz/browse_for_stats/people_and_communities/Well-being/social-connectedness/social-networks.aspx
- Van den Wollenberg, Arnold L. (1977). Redundancy analysis: An alternative for canonical correlation. *Psychometrika*, 42(2), 207-219.

Appendices

Appendix A

List of variables mentioned throughout the report, which are not necessarily found in Appendix D:

- **qSupportType**: The respondent identifies which types of support they required from the social network in relation to the most impactful life changing event.
- **qOneFamTalk/qFamTalk**: If the respondent has family out of the household, how frequent does he/she talk with them. A position generator.
- **qOneFamVideo/qFamVideo**: If the respondent has family out of the household, how frequent does he/she have video conversations with them. A position generator.
- **qOneFamPhone/qFamPhone**: If the respondent has family out of the household, how frequent does he/she phone them. A position generator.
- **qOneFamWrite/qFamWrite**: If the respondent has family out of the household, how frequent does he/she write to them. A position generator.
- **qFirstSupport**: The first person the respondent identified that gave them emotional/information and advice/financial/practical help during the most impactful life changing event.
- **LivesAlone**: The NZGSS 2014 identified that the respondent lives alone. A conditioning variable.
- **PrtnrInHH**: The NZGSS 2014 identified that the respondent lives with their spouse/partner. A conditioning variable.
- **ChldrInHH**: The NZGSS 2014 identified that the respondent lives with their child(ren). A conditioning variable.
- **AllFmlyInHH**: The NZGSS 2014 identified that the respondent lives with all of their family. A conditioning variable.
- **AllFrndInHH**: The NZGSS 2014 identified that the respondent lives with all of their friends. A conditioning variable.

Appendix B

Please find the file [Interaction](#) in the Appendices folder.

Appendix C

Please find the file [Anova Type-III](#) in the Appendices folder.

Appendix D

List of all non-conditioning variables in the final redundancy analysis object, where bold font indicates that a variable was present in the baseline model and italic font indicates that a variable was derived for this project.

- **DVAge**: The derived age of the respondent, generated by Statistics NZ.
- **DVSex**: The gender of the respondent.
- **DVEthTr_Euro**: Whether the respondent identified their ethnicity as European.
- **DVEthTR_Maori**: Whether the respondent identified their ethnicity as Maori.
- **DVEthTR_Pacific**: Whether the respondent identified their ethnicity as Pacific.
- **DVEthTR_Asian**: Whether the respondent identified their ethnicity as Asian.

- **DVEthTR_MELAA:** Whether the respondent identified their ethnicity as Middle Eastern / Latin American / African.
- **DVEthTR_Other:** Whether the respondent identified their ethnicity as an unlisted ethnic group.
- **DVEthTR_NS:** Whether the respondent failed to identify their ethnicity.
- **DVPersIncSourc_Benefit:** Whether the respondent has an income source which is classed as a benefit.
- **DVPersIncSourc_NotBenefit:** Whether the respondent has an income source which is not classed as a benefit.
- **SNCqFamilyNum:** How large is the respondent's family network. A partitioned continuous variable.
- **SNCqFamRelations_Close:** Whether the respondent has a spouse/partner, child, or sibling network tie within their family network.
- **SNCqSpouseTime:** How much subjective quality time does the respondent spend with their spouse or partner. A position generator.
- **SNCqChildTime:** How much subjective quality time does the respondent spend with their child. A position generator.
- **SNCqFamPhone:** If the respondent has family out of the household, how frequent does he/she phone them. A position generator.
- **SNCqFamTalk:** If the respondent has family out of the household, how frequent does he/she talk with them face to face. A position generator.
- **SNCqFriendsNum:** How large is the respondent's friend network. A partitioned continuous variable.
- **SNCqFrndHomophily_Var:** The variance between positions for the friend homophily questions in the NZGSS 2014. This is coded as a factor variable because it would be more interesting looking at the difference between groups (rather than as a continuous variable).
- **SNCqFrndHomophily_Max:** The highest answer in the position generator for the friend homophily questions.
- **SNCqHoodNum:** How many years has the respondent spent in their neighbourhood. A partitioned continuous variable.
- **SNCqHoodYears:** How large is the respondent's neighbourhood network. A partitioned continuous variable.
- **SNCqClubMembership:** Whether the respondent is part of a club.
- **SNCqHHGetAlong** How well does the respondent get along with their household, on a 0-10 rating scale. A partitioned continuous variable.
- **SNCqTimeLonely:** How much time does the respondent spend being alone. A position generator.
- **SONqNewsWho:** In the hypothetical situation of sharing good news, who would be the first person the respondent will share it with.
- **SONqNewsProximity:** The geographic proximity of the person the respondent identified.
- **SONqNewsAccess:** How easily accessible is the person the respondent identified.
- **SONqStayWho:** In the hypothetical situation of needing a place to stay away from the household, who would be the first person the respondent will stay with.

- **SONqStayProximity:** The geographic proximity of the person the respondent identified.
- **SONqStayAccess:** How easily accessible is the person the respondent identified.
- **EFFqHowSignificant:** How significant was the most impactful life changing event. A partitioned continuous variable.
- **EFFqNumOfLifeChngs:** The number of significant life changes in the last 12 months. A partitioned continuous variable.
- **EFFqPosOrNeg:** Whether the most impactful life changing event was positive or negative. A position generator.
- **EFFqAlterMatch_Score:** The similarity of a respondent's identified alters when compared to supportive people who provided immediate support during their most impactful life changing event. This measure counts the number of cases identified. A continuous variable.
- **EFFqSupType_Emoational:** Whether the respondent required emotional support during the most impactful life changing event.
- **EFFqSupType_InfoAdvice:** Whether the respondent required information and advice during the most impactful life changing event.
- **EFFqSupType_Financial:** Whether the respondent required financial support during the most impactful life changing event.
- **EFFqSupType_Practical:** Whether the respondent required practical support during the most impactful life changing event.
- **EFFqHelpProximity_Emoational_11:** The geographic proximity of an alter with regards to how effective their social network was for emotional support. A position generator.
- **EFFqHelpProximity_InfoAdvice_12:** The geographic proximity of an alter with regards to how effective their social network was for information and advice. A position generator.
- **EFFqHelpProximity_Financial_13:** The geographic proximity of an alter with regards to how effective their social network was for financial support. A position generator.
- **EFFqHelpProximity_Practical_14:** The geographic proximity of an alter with regards to how effective their social network was for practical support. A position generator.
- **EFFqSupHelpful_Emoational_11:** How helpful was the emotional support provided by an alter during the most impactful life changing event. A position generator.
- **EFFqSupHelpful_InfoAdvice_12:** How helpful was the information and advice provided by an alter during the most impactful life changing event. A position generator.
- **EFFqSupHelpful_Financial_13:** How helpful was the financial support provided by an alter during the most impactful life changing event. A position generator.
- **EFFqSupHelpful_Practical_14:** How helpful was the practical support provided by an alter during the most impactful life changing event. A position generator.

Appendix E

Please find the file [Biplot Arrow Distances \(Sorted\)](#) in the Appendices folder.

Appendix F

Please find the three R scripts I have prepared in the Appendices folder. [Analysis Step](#) and [Data Cleaning Step](#) are R scripts which are meant to be used together. [Export Script](#) is a R script which creates a dataset based on the NZGSS 2014. The resulting dataset contains the social network measures and demographic measures I have found to be statistically significant from Section 3.4.