SCHOOL OF ECONOMICS

UNIVERSITY OF NOTTINGHAM

ESSAYS ON WELL-BEING: A UK ANALYSIS

CHRYSANTHOS VASILEIOU

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

JULY 2023

Vasileiou, Chrysanthos (2023). Essays on Well-being: A UK analysis.
PhD Thesis, University of Nottingham

Supervised by Trudy Owens and Sarah Bridges

Nottingham, July 2023

ACKNOWLEDGMENTS

I am enormously grateful to my supervisors Dr Trudy Owens and Dr Sarah Bridges for their continuous support and advice. This would have not been possible without your invaluable contributions and guidance.

I want to express my deepest gratitude to my family Nikolas, Kyriaki, Vassilis, Marios, George, and Angeliki for being by my side in this difficult journey.

Finally, I want to thank the School of Economics at the University of Nottingham for giving me the opportunity to be part of such an amazing experience, and everyone at the department for making the path easier by being exceptionally helpful and welcoming.

CONTENTS

THESIS INTRODUCTION

As a preliminary note, readers should be aware that each chapter is written as a standalone paper on subjective well-being. As such, this thesis consists of the current introduction, followed by three separate papers of standard length with their own references and detailed appendices, and lastly a conclusion which summarizes the main findings. Some repetition of the relevant literature is unavoidable. This will be highlighted throughout wherever applicable.

Subjective, self-reported measures originate from responses to survey questions about aspects of life that are unobserved in their true form. These include life satisfaction and mental health, which constitute the main focus of the current thesis. Subjective measures offer a path for incorporating such concepts in quantitative analysis. Individuals are usually asked to rate facets of their lives relevant for each concept on an ordinal scale (known as a Likert scale).

The subjective well-being literature, which heavily uses the two aforementioned concepts to represent well-being, is characterised by two parallel strands of research. One strand investigates the validity of such measures in representing aspects such as mental health, and the other operates under the assumption that such measures are valid to examine which factors influence them. The current thesis adds to both strands of research. Using data from the UK Understanding Society household surveys it contributes to the evidence in support of the validity of subjective measures in chapter 1, offering an alternative approach to study the determinants of life satisfaction in chapter 2, and examining how the determination of mental health changed during the recent COVID-19 pandemic in chapter 3.

The motivation for chapter 1 stems from the fact that the support for subjective measures is not universal across the literature (see, Bertrand and Mullainathan, 2001; and Bond and Lang, 2019). For example, Bond and Lang (2019) criticize their use by demonstrating how some of the main results in well-being literature which use subjective measures can be reversed by monotonic transformations of well-being. Given that subjective measures are ordinal, results should be consistent to monotonic transformations since they do not influence the original order of well-being across individuals.

While the other two chapters assume subjective measures suffice for quantitative analysis, the aim of the first chapter is to test the validity of these measures, namely life satisfaction and mental health, in relation to observed biomarker and physiological data. Biomarker and physiological measurements are objective as they are recorded by a nurse or physician and can therefore be factual indicators of general health or well-being. The methodology used is known

as a regular vine copula. It provides measures of the association of subjective well-being with the biomarkers and physiological indicators which are invariant to monotonic transformations of well-being. The estimated model provides evidence in support of subjective measures accurately representing well-being when compared to objective health measures. For example, the biomarkers for glycated haemoglobin, diastolic blood pressure, dehydroepiandrosterone sulphate, forced vital capacity, albumin, and high-density lipoprotein cholesterol are significantly associated with subjective well-being.

There are studies which support the validity of subjective measures on the basis of strong associations with objective measures. Hamer and Chida (2011) provide evidence of an inverse association between life satisfaction and two inflammatory biomarkers, namely c-reactive protein, and fibrinogen. There are also findings such as the association of reported happiness with blood pressure, heart rate, and prefrontal brain activity (Alesina *et al.*, 2004). The current paper adds value by combining such measures in the same environment to examine them simultaneously. An additional contribution is to do this in conjunction with the copula-based methodology which offers invariance to monotonic transformations of the well-being scale in the characterisation of associations, among other advantages that revolve around modelling flexibility and are described in detail in chapter 1.

After establishing the validity of subjective measures, chapter 2 moves on to examine the determinants of life satisfaction through a machine learning technique which can be classified as a non-parametric approach. Most studies to date employ parametric methods such as linear regressions. Ferrer-i-Carbonell (2013), and Clark (2018) describe several important results which have emerged from the use of parametric approaches over the past four decades. Examples include the U-shaped relationship between well-being and age, the impact of social comparison concerns on well-being, and the adaptation of well-being to events that may have an initial impact on individual welfare.

The technique used in chapter 2 is the RE-EM tree by Sela and Simonoff (2012). This chapter complements the standard linear techniques to provide a well-rounded perspective. Tree-based methods require no *a priori* model structure or variable selection. When modelling life satisfaction this can prove useful as several non-linearities and interactions between explanatory variables, that would otherwise seem unlikely to be pre-specified, reveal themselves to the researcher. In order to facilitate comparison with standard techniques, the well-being structure suggested by the RE-EM tree is compared to a linear model. The

explanatory power of the two is comparable suggesting that the non-parametric estimation can offer useful insights and complement the traditional parametric approach. A predictive margins analysis is also carried out showing that the estimated RE-EM tree structure replicates many of the results in literature with regard to major determinants of well-being.

Despite being able to identify well-being determinants in a consistent manner across different methodologies, the stability of the well-being determination process is another aspect which needs examining. Chapter 3 focuses on the COVID-19 pandemic. Given the deterioration in the UK's average level of well-being as a result of the COVID-19 virus outbreak in 2020, it is then a question of whether the major determinants of well-being identified continue to influence well-being in the same way after the onset of the pandemic.

The COVID-19 pandemic impacted every aspect of life. It led to concerns about health, social restrictions, money worries, and job insecurity, all of which led to a large deterioration in mental health or well-being (see, Banks and Xu, 2020; Chandola *et al*., 2020; Daly *et al*., 2020; Banks *et al*., 2021). Understanding how the determinants of mental health change during a crisis allows targeted interventions at those in most need. Crises are often accompanied by structural breaks to other sectors of the economy, but to the best of our knowledge no study has investigated whether a crisis can also lead to a structural break in the determinants of mental health.

Understanding Society's COVID-19 web survey is used, in combination with the original survey used in the other two chapters, to examine whether there has been a structural break in the determinants of mental health (i.e., a change in the parameters of the regression models). The analysis focuses on the effect of seven variables for mental health during the pandemic: partnership status, feelings of loneliness, children, health status, employment status, hours worked, and income.

There is evidence of two structural breaks in the determinants of mental health. In line with expectations, the first occurs at the start of the pandemic, close to when the first lockdown was implemented in the UK. The second structural break occurs during the summer of 2020, shortly after many of the UK's COVID-19 restrictions had been eased, and heavily depends on the influence of feelings of loneliness on mental health.

In summary, the current thesis offers support for the use of subjective, self-reported well-being measures in chapter 1, and goes on to demonstrate how an alternative non-parametric estimation technique can offer useful insights in the determination of life satisfaction. The last

chapter presents the influence of the COVID-19 pandemic on well-being determination, where well-being is captured by a measure of mental health.

CHAPTER 1: COPULA-BASED CHARACTERISATION OF THE ASSOCIATION BETWEEN BIOMARKERS AND SELF-REPORTED WELL-BEING

Abstract: Many argue subjective self-reported well-being measures offer a plausible approach to incorporating unobserved concepts in quantitative analysis. The validity of such measures, however, has been questioned. This paper examines subjective self-reported well-being measures in relation to biomarker and physiological data in an attempt to investigate the legitimacy of self-reported measures in representing well-being. In doing so, it provides measures of the strength of association of subjective well-being with the biomarkers and physiological indicators which are invariant to monotonic transformations of latent well-being. The methodology used comes from copula theory, known as a regular vine copula. Two self-reported well-being measures from the UK Understanding Society data set are examined: life satisfaction and the combination of the 12 questions of the General Health Questionnaire. The estimated model suggests evidence in favour of self-reported measures in accurately representing well-being when compared to more objective health measures. For instance, the biomarkers for glycated haemoglobin, diastolic blood pressure, dehydroepiandrosterone sulphate, forced vital capacity, albumin, and high-density lipoprotein cholesterol are significantly associated with self-reported well-being.

1. INTRODUCTION

Subjective, self-reported measures refer to responses to survey questions which may be used to elicit information about latent variables associated with different aspects of life, such as life satisfaction, happiness, anxiety, and mental health. Such measures are usually recorded on Likert scales resulting in variables of ordinal nature. Many argue they offer a plausible approach to incorporating unobserved concepts in quantitative analysis. They are used extensively in literature to approximate aspects such as happiness (Alesina *et al*., 2004; Clark *et al*., 2008), pain (Blanchflower and Oswald, 2019), and cultural traits (Alesina and Juliano, 2015). However, the support for such measures is not unanimous which provides the motivation for this paper. While the other two chapters assume that self-reported measures suffice for quantitative analysis, the aim of this chapter is to test the validity of these self-reported measures in relation to observed biomarker and physiological data. Biomarker and physiological measurements are objective in that they are recorded by a nurse or physician and can act as indicators of general health or well-being.

A significant part of the well-being literature deals with the accuracy, and thus usefulness, of subjective measures in capturing what they aim to record (see for example, Bertrand and Mullainathan, 2001; Bond and Lang, 2019; and Oswald and Wu, 2010). Relevant measures are frequently treated as ordered responses on an interval scale or analysed under the assumption that the latent variable distribution is either normal or logistic (Bond and Lang, 2019). There is evidence supporting that the two substitute assumptions yield qualitatively similar results (Ferrer-i-Carbonell and Frijters, 2004). As such, the argument can be made that the choice of methodological approach does not significantly influence the inferences made with respect to well-being.

However, one of the main concerns raised with regard to subjective, self-reported measures capturing well-being is that, without imposing strong auxiliary assumptions, it is hard to support the comparison of groups of individuals based on the estimated mean values of their true underlying well-being distributions by using survey data. Bond and Lang (2019) operate in a context in which happiness[1] is considered similar to the notion of utility, and therefore there can be infinite candidates for the true underlying happiness distribution that can preserve

[1] The authors are using the notion of happiness, but their conclusions can be extended to incorporate any other unobserved variable for which the distribution is approximated by subjective responses to survey questions recorded on some form of ordinal scale.

the choices observed to be made by individuals on the ordinal scale provided to them[2]. As such, the only way to make sure that the mean ranking of groups remains the same for all possibilities is to establish that the happiness distribution of one group first order stochastically dominates (FOSD) the other. The authors propose that it is highly unlikely for the conditions of non-parametric identification of stochastic dominance to be met in practice (e.g. groups cannot be ranked in terms of FOSD if both have observed responses in the highest and lowest categories of the survey's ordinal scale). When it comes to parametric identification, the authors suggest that it is almost impossible to establish stochastic dominance in the case that the underlying distributions of the groups come from the same unbounded location-scale family. The reason behind this is that arbitrary strict monotonic transformations of the scale can reverse the mean ranking of groups. The only chance of identifying stochastic dominance is in the unlikely case of the equality of variances of the happiness distributions between the groups under consideration. Bond and Lang (2019) demonstrate how some of the main results in happiness literature can be reversed by monotonic transformations of latent happiness. The main takeaway is that the assumption about the latent well-being distribution matters.

The aim of this paper is to examine the informational content of self-reported well-being in relation to biomarker and physiological data. Examining the informational content in this paper refers to studying the bilateral associations of well-being with each biomarker, while controlling for the rest of the biomarkers, in an attempt to provide support for the validity of the subjective well-being data. As such, the validity of subjective well-being is established on the grounds of its association with the biological well-being of individuals which is represented by the basic biomarkers used in this paper. Meaningful evidence in support of subjective well-being should suggest that higher self-reported well-being is associated with higher biological well-being.

A subsequent objective is to provide a measure of the strength of association with biomarkers and physiological measures that is invariant to strictly monotonic transformations of latent well-being. Loosely speaking, we operate under the assumption of an unobserved absolute value of well-being associated with each level of the ordinal well-being measure used in the current study. The measures of association derived should be unaffected by the assumption for the unobserved absolute well-being values associated with each level of the ordinal scale as

[2] This is similar to the idea that any strict monotonic transformation of a utility function represents the same preference ordering. In this case the ordering is concerned with states of happiness. As such, there is an infinite number of arbitrary cardinalizations that could represent the self-reported data equally well.

long as the original well-being ranking is preserved. This latter objective is linked to the issue proposed by Bond and Lang (2019). The invariance to monotonic transformations does not provide support for the validity of self-reported well-being per se, but rather means that the characterisation of any significant association of well-being which is unveiled remains unaltered to any assumption with regard to the shape of the true underling well-being distribution.

The methodology used comes from copula theory. Copula theory provides the tools for the examination of the dependence between random variables in a manner which allows studying the dependence structure separately from each of the univariate marginal distributions of the variables. A copula function can be used to represent the dependence structure. Under certain conditions, any strict monotonic transformation of the variables considered will not alter the dependence structure as represented by the copula. In that sense, copulas allow modelling of the association between variables of interest in a 'margin-free' way.

Two self-reported well-being measures are used in the analysis. They come from a sample consisting of pooled cross-sections of individuals generated by merging two sets of observations, from waves 2 and 3 of the UK Understanding Society data set. The first self-reported variable represents overall life satisfaction recorded on a 7-point Likert scale. The second self-reported variable combines the 12 questions of the General Health Questionnaire (each recorded on a 4-point Likert scale) to a one-dimensional measure with 37 levels (GHQ). The life satisfaction variable is extensively used in literature to approximate well-being (see Becchetti *et al.,* 2013; Boyce *et al.,* 2010; Di Tella *et al.,* 2010; and Ferrer-i-Carbonell, 2005). GHQ is also a measure that is used to capture well-being, in a sense which is closer to the mental health component of well-being (see Brown *et al.,* 2015; Clark and Oswald, 2002; and Wood *et al,* 2012). By combining these two widely used measures, the aim is to represent subjective, self-reported well-being in a manner which is as inclusive as possible. More details regarding the way in which the measures are used are provided in subsections 5.1 and 5.2.

One possible approach in the attempt to support the usefulness of subjective well-being measures is to establish their relationship with more palpable quantities. Biomarkers can play the role of such quantities. Biomarkers represent objectively measured characteristics that can be used as indicators of the biological well-being of individuals. The set of biomarkers chosen for this paper represents multiple functions. It includes markers for adiposity, blood pressure, cholesterol levels, lung function, inflammation, blood sugar levels, liver function, and a steroid

hormone (Davillas and Pudney, 2020). Along with the use of biomarker and physiological data comes the last objective of the paper which is to characterise the joint distribution of biomarker, physiological, and self-reported data. Bivariate associations could be modelled directly (i.e. unconditional relationships). However, it is more useful to model bivariate associations in an environment in which the rest of the variables are controlled for to avoid any overlap in the information provided by the different biomarkers.

Various studies which attempt to support the validity of self-reported measures based on the ground of strong associations with less subjective measures already exist. These are studies attempting to establish the association of subjective well-being measures to different aspects in the life of an individual which are more 'tangible'. Oswald and Wu (2010) show how life satisfaction is strongly correlated across geographical areas with a quality-of-life measure accounting for regional attributes such as sunshine, temperature, crime, etc. Hamer and Chida (2011) demonstrate a linear inverse association between life satisfaction and two inflammatory biomarkers, namely c-reactive protein, and fibrinogen. Alesina *et al*. (2004) present findings such as the association of reported happiness to measures including blood pressure, heart rate, and prefrontal brain activity. The current paper adds value by attempting to combine such measures in the same environment to analyse them simultaneously. An additional contribution to the literature is to do this in conjunction with a copula-based methodology, outlined in detail in Appendix A and Appendix B. Copulas are used as representations of the dependence between different measures and are invariant to certain assumptions that can be made regarding the true well-being distribution.

The next section of the paper provides a literature review, followed by an outline of the basic background on copula theory. A data section is then provided, followed by a section which outlines the subjective, self-reported well-being measure used for the analysis. The penultimate section deals with the results of the analysis, followed by a concluding, summative section.

2. LITERATURE REVIEW

2.1 Subjective well-being

A major part of the literature assumes that subjective well-being measures are adequate proxies for the underlying level of well-being and go on to investigate the structural form of their data-generating processes (see for example, Gerdtham and Johannesson, 2001; Clark and Oswald, 2002; and Boyce *et al*., 2010). For example, for life satisfaction, it is assumed that the self-reported measure used to capture life satisfaction does indeed encapsulate the actual, unobserved value of the variable for each individual in a satisfactory manner. Based on this assumption, focus can then be given to understanding the determinants of life satisfaction.

Ferrer-i-Carbonell (2013), and Clark (2018) offer comprehensive reviews of the findings associated with this part of the literature. Several important results have emerged such as the U-shaped association of well-being with age[3], the significant impact of social comparison on well-being[4], as well as the adaptation of well-being across time to events that initially may have had a significant impact on individual welfare[5]. The findings presented in the two reviews include the association of a higher level of well-being with higher levels of income[6] and health, the negative relationship between well-being and unemployment (as opposed to being employed), and the negative association between well-being and being single (as opposed to being married). Such findings date back to one of the first studies by Gerdtham and Johannesson (2001) who study the concept of happiness.

Various concepts are used across the different studies, including happiness and life satisfaction. Clark (2015) considers three main types which include life satisfaction, affect, and eudaimonia[7]. He finds that they are significantly correlated with each other, as well as being associated with certain explanatory variables such as education, marital status, and the natural logarithm of income in approximately the same way.

Regardless of the concept under consideration, as long as the measures used are self-reported by individuals and recorded on an ordinal scale, the concerns about their accuracy reported in the previous section remain. Several studies have attempted to provide empirical support for the validity of the measures. A common approach is to associate the subjective measures with

[3] See for example Blanchflower and Oswald (2008b), Glenn (2009), and Frijters and Beatton (2012).
[4] See for example Clark and Oswald (2002), Ferrer-i-Carbonell (2005), and Becchetti *et al.* (2013).
[5] See for example Ferrer-i-Carbonell and Van Praag (2008), and Oswald and Powdthavee (2008).
[6] This is true in the case that cross-sectional data is used.
[7] Affect has to do with an instantaneous judgment of how an individual is feeling. Eudaimonia is a concept dealing with an individual achieving potential in various aspects of life.

variables which capture the well-being of individuals, and at the same time attempt to minimize the subjective component involved in asking individuals to assess their own well-being. Such studies which report on the associations of the measures with biomarkers will be reported in the next subsection. In the current subsection, studies which report on the associations with objective measures of well-being outside biomarker variables are reported.

Oswald and Wu (2010) support the validity of self-reported measures by showing how they are strongly correlated with measures constructed based on non-subjective data across geographical areas for a compensating-differentials approach. The authors use data from the U.S. Behavioral Risk Factor Surveillance System to construct regression-adjusted[8] life satisfaction estimates for 50 U.S. states. These estimates, representing the subjective version of measuring life satisfaction, are found to have a strong association with an objective quality-of-life ranking for the states, constructed by Gabriel *et al.* (2003) based on indicators measuring aspects such as sunshine, temperature, violent crime, air quality, student-teacher ratio, taxes, and many other features of life.

Another objective dimension used to support the validity of self-reported measures is smiling. Intuitively, happier individuals should smile more. Ekman *et al.* (1990) provide evidence for the higher frequency of the Duchenne smile[9] during pleasant experiences as opposed to unpleasant experiences. The authors attempt to replicate pleasant and unpleasant experiences by using films that individuals are asked to watch. The subjective reports on positive emotions from the subjects are also recorded. They appear to be positively related with the frequency of the Duchenne smile.

A concept closely related to the one of biomarkers is the study of brain activity. It is related in the sense that it too represents an attempt to elicit information about well-being directly from the horse's mouth. Advances in technology make it possible to detect brain activity by using an electroencephalogram (EEG). Sutton and Davidson (1997) find that asymmetry between right and left prefrontal activation of the brain is related to the relative self-reported strength of the behavioural approach (BAS) and inhibition (BIS) systems. Based on the scale on which they are recorded by the authors, the BAS and BIS represent hypothetical systems which

[8] Adjusted by controlling, among other things, for income, age, gender, ethnicity, education, marital, and employment status.

[9] The Duchenne smile is used as an indication of true enjoyment as opposed to other types of smiles or facial expressions in general. The pattern of facial muscle activations is used to identify the Duchenne smile.

account for the tendency of an individual to experience intense positive and negative affect respectively.

An interesting approach to study the validity of self-reported measures is by considering how they are related to measures that capture the well-being of an individual but are reported by people close to the particular individual. Sandvik *et al*. (1993) collect reports by friends and family members for 136 university students with regard to their well-being. The authors use both friends and family for their experiment as they want to capture well-being in all possible situations that the students might find themselves in. There is consistency in the reporting of students' well-being across the two groups of friends and family members. In support of the validity of self-reported measures, the well-being of students appears to be consistent across self-reported measures and the measures provided from friends and family.

Self-reported measures can also be studied at a macroeconomic level. A national index of well-being can be calculated by averaging across individuals in the sample from a particular country. In doing so, Di Tella *et al*. (2003) regress national suicide rates on well-being. They operate in a panel context incorporating fixed effects at the country level in their estimation. Based on their results, a higher level of national self-reported well-being is associated with a lower national suicide rate.

2.2 Biomarkers

A complete list of the biomarkers used in the present study is offered in subsection 4.2. The particular choice of biomarkers for this paper is also mostly in agreement with studies such as Davillas and Pudney (2017, 2020) who use the same data set to study the concordance between the health states of partners who are in a marital or cohabitating relationship; and the determinants of demand for primary and secondary public healthcare services which stem from individual-level characteristics.

Several of the biomarkers used in the present study are examined separately throughout the well-being literature. Hamer and Chida (2011) use the Scottish Health Survey of 2008 to study the relationship between life satisfaction[10] and inflammatory biomarkers. The authors consider c-reactive protein and fibrinogen which are collected through blood samples. There is evidence

[10] Self-reports are recorded on a 10-point Likert scale.

for the case of a linear negative association between life satisfaction and each of the two inflammatory biomarkers[11].

Blood pressure is a biomarker used in many of the studies dealing with well-being. Blanchflower and Oswald (2008a) attempt to provide credibility for measures such as self-reported happiness and life satisfaction. The authors use a self-reported measure of blood pressure as a proxy for the existence of hypertension[12] with the presumption that individuals reporting high blood pressure transmit the information given to them by qualified doctors. Based on data from 16 countries, the study suggests a systematic inverse relationship between blood pressure and self-reported well-being. Countries which exhibit a higher level of well-being tend to also report lower levels of hypertension. Romero Martinez *et al*. (2010) examine the association between life satisfaction and blood pressure in adolescents. The authors use a logistic regression to conclude that high blood pressure is associated with a low level of life satisfaction, especially for male adolescents. However, the study uses binary variables as measures for both blood pressure and well-being. Szabo *et al*. (2020) examine the relationship between life satisfaction and blood pressure in a sample of 68 adults. They identify a significant negative association between systolic blood pressure and life satisfaction, whereas no significant relationship is found between life satisfaction and diastolic blood pressure.

Along with blood pressure, the level of cholesterol can also be an important indicator of cardiovascular health and thus of the general well-being for an individual. Radler *et al*. (2018) examine the association between well-being and high-density lipoprotein cholesterol[13] (*hdl*). The authors consider well-being as represented by eudaimonia[14]. They find that a higher level of well-being is associated with a higher level of *hdl*. The level of cholesterol is even examined in relation to depression and suicides. In a review of the literature, Manfredini *et al*. (2000) consider the possible link between the reduction[15] of serum cholesterol[16] and the increase in violent deaths and suicide. The authors conclude that there is not enough evidence against the prescription of cholesterol-lowering drugs.

[11] The authors control for other variables including age, sex, education, smoking, body mass index, and depressive symptoms.
[12] Hypertension refers to blood pressure which is higher than what it should normally be.
[13] The 'good' type of cholesterol.
[14] Individuals report on questions regarding autonomy, environmental mastery, personal growth, positive relations with others, purpose in life, and self-acceptance.
[15] Such as the one during a treatment against coronary heart disease.
[16] The measurement of serum cholesterol provides an indication for the level of elements in the blood including high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, and triglycerides.

Glycated haemoglobin (*hba1c*) is a biomarker used to track the sugar level in the blood. Tsenkova *et al*. (2008) use a sample of 97 elderly women to demonstrate that higher levels of positive affect are associated with lower levels of *hba1c*. The authors use regression analysis in which the dependent variable is the level of *hba1c*[17]. Poole *et al*. (2019) are also concerned with the elderly by using the English Longitudinal Study of Ageing for their study. They provide evidence for a negative association between subjective well-being and *hba1c* through regression analysis which also controls for depressive symptoms, age, and sex. It should be noted that *hba1c* is measured 8 years after the subjective well-being indicator[18] is recorded.

Dehydroepiandrosterone sulphate (*dheas*) is a steroid hormone. Higher levels of *dheas* are associated with better health. Wong *et al*. (2011) present a negative association[19] between *dheas* and the level of depressive symptoms[20] in a sample of elderly Chinese men. Valtysdottir *et al*. (2003) study the relationship between *dheas* and mental well-being in a sample of female patients with primary Sjögren's syndrome and find a significant positive association. Well-being is self-reported through the Psychological General Well-Being Index which captures dimensions such as anxiety, depression, positive well-being, self-control, general health, and vitality. The possibility of cultural variations in the association between *dheas* and well-being is also examined in the literature. Yoo *et al*. (2016) consider how the relationship between positive affect and the biomarker varies with social connectedness across two countries, namely Japan and the U.S. Positive affect is recorded using a 10-item measure which requires individuals to state the frequency of experiencing a list of feelings in the 30 days prior to the interview. The feelings include cheerful, in good spirits, extremely happy, calm and peaceful, satisfied, full of life, enthusiastic, attentive, active, and proud. The authors find that a high level of positive affect combined with low social connectedness is linked with a low level of *dheas* in Japan. For the U.S. there appears to be no apparent relationship between positive affect and *dheas*.

Studies have also considered the relationship between self-reported well-being and albumin (*alb*) used to capture liver function, or the association between self-reported well-being and forced vital capacity (*htfvc*) used to approximate respiratory function. Goracci *et al*. (2008)

[17] The authors control for sociodemographic characteristics and health condition of the individuals.
[18] Subjective well-being is captured by the CASP-19 score which considers the domains of control, autonomy, self-realization, and pleasure.
[19] Marginally significant at the 10% significance level.
[20] Recorded through the Chinese Geriatric Depression scale.

consider 80 individuals with sarcoidosis[21] and find that *htfvc* is positively related with several dimensions of self-reported well-being. The authors use the Quality of Life Enjoyment and Satisfaction Questionnaire to capture the self-reported well-being of individuals across several dimensions. Individuals are asked to report on feelings, work, social relations, physical health/activities, household duties, leisure time activities, school/course work, and general activities. As far as *alb* is concerned, Schenk *et al*. (2018) use the biomarker to construct a composite measure of allostatic load[22]. For the constructed measure, a lower *alb* level contributes to an elevated allostatic load. Using a regression analysis which adjusts for age, sex, and negative affect the authors report an inverse association between a self-reported measure of positive affect and the level of allostatic load. The levels of positive and negative affect are recorded based on the Positive and Negative Affect Schedule (PANAS) which requires the individuals to report the frequency of experiencing certain feelings in the four weeks prior to the interview. As far as the positive affect component is concerned, feelings include enthusiastic, proud, alert, inspired, interested, excited, strong, determined, attentive, and active. For the negative affect component feelings include distressed, afraid, jittery, upset, guilty, scared, nervous, ashamed, hostile, and irritable. Literature also looks at this type of relationships in very specific groups of individuals. Prinsloo *et al*. (2015) study the relationship between positive affect and *alb*, as well as between depressive symptoms and *alb* in a sample of patients with metastatic renal cell carcinoma[23]. The Centers for Epidemiologic Studies-Depression (CES-D) assessment is the self-reported measure used to evaluate the level of depressive symptoms. A subscale of the CES-D is used to capture positive affect. They find a significant positive relationship with the level of positive affect, and a significant negative one with the level of depressive symptoms.

2.3 Copulas

The current paper uses a particular type of copula known as a regular vine copula. Copulas are extensively applied in bivariate analysis. A regular vine copula attempts to translate the benefits of the application of bivariate copulas to higher dimensions.

When it comes to bivariate analysis, Quinn (2007b) uses copula theory to measure the relationship between health and income in the European Community Household Panel Survey. Health is recorded through an ordinal self-reported measure. As such, the main advantage of

[21] An inflammatory disease that can affect several organs.
[22] The impact of chronic stress on the body.
[23] Kidney cancer that spreads to other parts of the body.

using a copula as a measure of association between the two variables is the invariance of the copula measure to any cardinalizations applied to the health measure. This benefit is also applicable to the current paper. Quinn (2007a) promotes the application of copulas to health economics in general due to their useful properties and versatility, mainly the fact that the dependence structure and the distribution functions of each random variable can be modelled independently. The author demonstrates their usefulness in areas such as health insurance and health care utilisation.

Some attempts are made to incorporate copula theory in the analysis of well-being and biomarkers. For example, Decancq (2013) uses copulas to model the dependence between three dimensions of well-being, namely income, health, and education. In doing so, the author proposes the use of a copula-based framework over standard measures such as the Human Development Index (HDI) to capture the levels of well-being in different societies. The reasoning behind the proposal lies in the fact that a copula-based framework accommodates for the dependence between the dimensions of well-being, whereas measures such as HDI do not as they rely on dimension-specific summary statistics. Societies can be very different in terms of the dependence between the well-being dimensions even if they might look similar in each dimension independently.

As far as biomarker research is concerned, Hutson *et al*. (2015) propose a copula-based framework for capturing the dependence between different biomarkers. The suggested multivariate epsilon–skew–normal distribution nests the standard multivariate normal model. The authors demonstrate the usefulness of the proposed model by examining the association between salivary biomarkers. They find that the joint probability densities of various pairs of biomarkers deviate substantially from the bivariate normal one. Their proposed model can capture non-linear dependencies between variables for which the common multivariate normal model would be a poor choice. Copulas are also used in studies which deal with medical conditions. For example, Kim *et al*. (2020) use a copula approach to generate predictions for the occurrence of dengue hemorrhagic fever in dengue infected individuals based on a set of biomarkers including indicators of weight, age, and lymphocytes[24]. Rakonczai *et al*. (2015) use copula models to examine the association between biomarkers linked to rheumatoid arthritis.

The finance literature provides some examples of applying regular vines. Zhang *et al*. (2018) use a regular vine copula to jointly model the Financial Stress Indices from 11 European

[24] A type of white blood cell.

countries[25]. The aim of the authors is to study the tail dependence between the different indices. This also constitutes one of the main advantages of using copula theory in that it offers great flexibility in terms of the properties of the distributions that can be selected. The authors note how the regular vine copula results highlight the complexity of the dependence structure of the various indices. This shows the adequacy of regular vine copulas in high-dimensional problems as in the present paper. Another example is the study of Mejdoub and Ben Arab (2018) who demonstrate the application of drawable vine copulas, a subclass of regular vine copulas, on the modelling of non-life insurance risks. The authors use the drawable vine copula for risk aggregation to capture the dependence structure between the different business lines of an insurance company[26].

Despite some of the literature on well-being and biomarkers acknowledging the value of using copula theory, to the best of the author's knowledge this is the first time that a regular vine copula is used in such a context.

[25] The authors incorporate a preliminary step in which they model each of the series using ARMA-GARCH filters.
[26] The authors use a simulation-based estimation of the Value at Risk (VaR) and the Tail Value at Risk (TVaR) which stems from the drawable vine copula estimation.

3. METHODOLOGY

A copula-based approach is used to model the association between subjective well-being and variables capturing biological well-being. In particular, a regular vine copula model is used. The current section provides a brief summary of the relevant methodology. Detailed descriptions of copula theory, and the regular vine copula along with its estimation procedure are provided in Appendix A and Appendix B respectively.

Ideally, one would like to know the joint distribution function of the variables of interest. The joint distribution function is made up of the individual variables' distributions along with some dependence structure which captures the association between the variables. In copula theory, this dependence structure is represented by a copula.

Sklar's theorem (1959) lies at the heart of copula theory, and states that for a set of d variables $\boldsymbol{X}$ with a joint distribution function $F_{\boldsymbol{X}}$ and marginal distribution functions F_i for $i \in \{1, \dots, d\}$, there exists a copula $C_{\boldsymbol{X}}$ such that:

$$F_{\boldsymbol{X}}(x_1, \dots, x_d) = C_{\boldsymbol{X}}\big(F_1(x_1), \dots, F_d(x_d)\big).$$

If all the variables are continuous then the copula implied by Sklar's theorem is unique. In addition, the copula remains unchanged in the case that any strict monotonic transformation is applied to any of the variables. This is linked to the issue reported in the introductory section with respect to the true unobserved well-being distribution. Invariance across many alternative cardinalizations of the reported scale is a desirable feature for dependence characterisation.

Parametric copulas offer great flexibility for the bivariate case in terms of the different features which they can encapsulate (e.g. asymmetry and tail dependence). However, as the size of the set of variables considered increases the level of flexibility is not the same. Vine copulas circumvent this issue by exploiting the large number of options for the bivariate case in a multivariate setting by modelling the associations between pairs of variables in the set of interest. Repeated conditioning is used to represent any multivariate density function as the product of unconditional and conditional bivariate densities. Bivariate copula densities are used to model the unconditional or conditional bilateral associations between variables. Each of these bivariate densities can be selected from the large set of parametric choices offered for the bivariate case, thus allowing to translate the flexibility offered in the bivariate setting to a higher number of dimensions.

The copula selection step is part of the so-called Dißmann's algorithm. The algorithm proceeds sequentially, selecting the order of conditioning used to decompose the multivariate density into bivariate (conditional) densities. The selection and estimation of the relevant bivariate copulas follows. The order of conditioning selected is such that the 'strongest' (conditional) associations between variables are modelled first. An independence test precedes the copula selection step to determine whether the two variables under consideration are (conditionally) independent.

4. DATA

4.1 Understanding Society

The main data set used to estimate the vine copula comes from waves 2 and 3 of the UK Household Longitudinal Study (UKHLS), which is also known as Understanding Society, in combination with Understanding Society's study on health and biomarkers. Understanding Society is a multi-purpose nationally representative survey of British households, including extensive socio-economic and psychological modules. Waves 2 and 3 were collected from 2010 up to, and including, 2012. These are the only two waves for which the biomarker and physiological data is recorded. The participants of the health and biomarkers study can be tracked back to the original study, thus allowing the two data sets to be merged.

The working sample consists of 8,154 individuals. It is constructed based on the condition that for each individual there are no missing values with regard to the self-reported well-being, biomarker, and physiological variables incorporated in the analysis. The particular sample used includes individuals from England and Wales. An extensive presentation of the biomarker and physiological data used is given in subsection 4.2. For the self-reported well-being variables, there are detailed definitions in subsection 5.1.

4.2 Biomarker and physiological data

Biomarkers Definitions Working Group (2001) defines a biological marker (biomarker) as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention".

As part of the Understanding Society survey, trained nurses visited the houses of the survey respondents approximately five months after the original survey (Benzeval *et al.*, 2014). The nurse's visit took place during wave 3 for those individuals who were part of the British Household Panel Survey (BHPS), the precursor of the Understanding Society survey. The relevant visit took place during wave 2 for the individuals of the non-BHPS sample[27]. As part of the visit, the nurses collected a set of biomarkers for each individual, either based on direct measurements during the visit, or through non-fasted blood samples.

The choice of the biomarkers incorporated in this study is based on an attempt to capture the overall state of health (or well-being) for each individual. As such, the set of biomarkers used

[27] Individuals eligible for nurse visits were those aged 16 or over who lived in England, Scotland, or Wales. For women, not being pregnant was a condition too. For the blood samples, individuals with no clotting or bleeding disorders were considered. Those with a history of fits were also not eligible. More details can be found on https://www.understandingsociety.ac.uk/documentation/health-assessment/user-guide.

aims to capture different dimensions of health. This is an approach considered by Davillas and Pudney (2020) as well, when trying to derive a composite index for health.

Variables capturing the height (*height*) and weight (*weight*) of individuals are used to capture the physical attributes of each individual, including characteristics such as adiposity[28]. In addition, a variable accounting for the age (*age*) of each individual is included in the analysis.

Measures of both systolic (*sys*) and diastolic (*dias*) blood pressure are used to capture the level of cardiovascular health[29]. For the same reason, the high-density lipoprotein cholesterol (*hdl*) measure is also included. Systolic (diastolic) blood pressure represents the maximum (minimum) pressure of the cardiac cycle (McFall *et al.*,2014). Increased blood pressure is associated with higher risk of cardiovascular disease. High-density lipoprotein cholesterol is considered as the 'good cholesterol' since *hdl* is involved in the transfer of cholesterol to the liver where it is broken down (Benzeval *et al.*, 2014). As such, low levels of *hdl* are associated with increased risk of cardiovascular disease.

Respiratory function is approximated by forced vital capacity (*htfvc*). Forced vital capacity measures the total amount of air which can be forcibly blown out after full inspiration (McFall *et al.*,2014). Higher values for *htfvc* are associated with better respiratory functioning.

C-reactive protein (*hscrp*) is included as a biomarker of inflammatory load. A higher quantity of *hscrp* in the blood is associated with a response of the body to inflammation (Benzeval *et al.*, 2014). Values greater than 3 mg/L represent systemic inflammation[30].

Glycated haemoglobin (*hba1c*) is included as a measure of the sugar level in the blood. It is an indicator of diabetes risk (World Health Organisation, 2011). Higher values can be used to diagnose diabetes, or poor management of diabetes (Benzeval *et al.*, 2014).

The biomarker albumin (*alb*) is used to capture liver function. Low levels of albumin are associated with possible loss of liver function (Benzeval *et al.*, 2014).

Lastly, dehydroepiandrosterone sulphate (*dheas*) is included which is a steroid hormone. Low levels of *dheas* are associated with cardiovascular disease (Barrett-Connor *et al.*, 1986) and

[28] The state of being obese.
[29] Three readings were taken for each of the two measures of blood pressure. The average of the readings is used for each measure in this study.
[30] Individuals with a recorded *hscrp* value strictly greater than 10 mg/L are excluded from the analysis as these values usually reflect recent infection (Pearson *et al.*, 2003).

higher levels are associated with better health (Benzeval *et al.*, 2014). *Table 1* provides summary statistics for each of the variables presented in this subsection.

Table 1: Summary statistics of biomarkers.

Biomarker (units of measurement)	**Sample mean**	**Sample standard deviation**
Height (cm)	167.918	9.512
Weight (kg)	78.380	15.844
Forced vital capacity (L)	3.879	1.085
Albumin (g/L)	47.018	2.806
Dehydroepiandrosterone sulphate ($\mu mlol/L$)	4.691	3.208
Glycated haemoglobin ($mmol/mol$)	36.879	7.616
High-density lipoprotein cholesterol ($mmlol/L$)	1.562	0.457
C-reactive protein (mg/L)	2.005	1.956
Age ($years$)	51.480	16.544
Systolic blood pressure ($mmHg$)	126.780	16.364
Diastolic blood pressure ($mmHg$)	73.486	10.569

Notes: Sample consists of 8,154 observations.

5. LIFE SATISFACTION AND GHQ

5.1 Definitions

The primary self-reported well-being variable in the subsequent analysis is life satisfaction. The variable used comes from individuals being asked to assess their life overall as a response to the question:

"Here are some questions about how you feel about your life. Please choose the number which you feel best describes how dissatisfied or satisfied you are with the following aspects of your current situation: Your life overall."[31]

The responses to this question are recorded on a 7-point Likert scale ranging from 1 being *"Completely dissatisfied"* to 7 being *"Completely satisfied"*.

The secondary self-reported well-being variable is a measure constructed based on responses to 12 questions of the General Health Questionnaire (GHQ). 12 questions are presented to individuals in the following manner:

"The next questions are about how you have been feeling over the last few weeks.

1. *Have you recently been able to concentrate on whatever you are doing?*
2. *Have you recently lost much sleep over worry?*
3. *Have you recently felt that you were playing a useful part in things?*
4. *Have you recently felt capable of making decisions about things?*
5. *Have you recently felt constantly under strain?*
6. *Have you recently felt you could not overcome your difficulties?*
7. *Have you recently been able to enjoy your normal day-to-day activities?*
8. *Have you recently been able to face up to problems?*
9. *Have you recently been feeling unhappy or depressed?*
10. *Have you recently been losing confidence in yourself?*
11. *Have you recently been thinking of yourself as a worthless person?*
12. *Have you recently been feeling reasonably happy, all things considered?"*

The responses to these questions are recorded on a 4-point Likert scale ranging from 1 being either *"More so than usual"* or *"Not at all"* to 4 being *"Much less than usual"* or *"Much more than usual"* depending on whether the question is of a 'positive' nature (e.g. question *1*) or a

[31] Questionnaires available on https://www.understandingsociety.ac.uk/documentation/mainstage/questionnaires.

'negative' nature (e.g. question *2*) respectively. The Understanding Society study summarizes the responses to these questions into a single number for each individual. The questions are converted to a single scale by recoding so that the Likert scale runs from 1 to 3 rather than 1 to 4, and afterwards summing, resulting in a measure which runs from 0 for "*Least distressed"* to 36 for *"Most distressed"*. For the purposes of the current study, the GHQ measure is reversed so that a higher number represents a lower level of distress. Therefore, for both measures, a higher integer value accounts for a higher level of well-being.

5.2 Composite self-reported well-being

For the purposes of this study, a composite measure of self-reported well-being is constructed. The life satisfaction measure is used as the primary signal of subjective well-being. The GHQ measure is used as the secondary indicator. It can be used to break the ties in subjective well-being between individuals reporting the same level of life satisfaction, generating a more refined ranking of the individuals in terms of subjective well-being. This is an approach used by Decancq (2013). In copula analysis, it is useful to have variables which are close to continuous. A more detailed discussion on this issue is provided in the subsection on robustness checks.

Given the ordinal nature of both variables, it is the case that ties in terms of self-reported well-being inherently exist between individuals in the sample as a limited number of possible responses are available[32]. Two or more observations are tied with respect to a particular variable when the recorded value of that variable for each observation is identical. The composite measure is a construct which attempts to (partially) break the ties in the primary self-reported well-being variable by using the secondary self-reported well-being variable in a lexicographic way. Lexicographic in this case translates to ranking two individuals in terms of well-being based on their responses to the life satisfaction question. A higher value for life satisfaction implies a higher value in the constructed measure. If they have identical responses to how they evaluate their satisfaction with life, they are ranked based on their GHQ score. Therefore, for individuals reporting identical life satisfaction, a higher value of GHQ implies a higher value in the constructed measure. Otherwise, they are assumed to have an equal level of self-reported well-being in the constructed measure.

[32] The sample size is large enough to ensure this.

More formally, for any individual $i \in \{1, \dots, n\}$ in the sample, the life satisfaction variable takes a value $ls_i \in \{1, \dots, 7\}$, the GHQ score takes a value $ghq_i \in \{0, \dots, 36\}$[33], and the constructed variable takes a value $r_i \in \mathbb{R}$. For any two individuals $i, j \in \{1, \dots, n\}$ in the sample where $i \neq j$ the vector $(r_i, r_j) \in \mathbb{R}^2$ is such that:

$ls_i > ls_j \Rightarrow r_i > r_j$,

$(ls_i = ls_j) \wedge (ghq_i > ghq_j) \Rightarrow r_i > r_j$,

and $(ls_i = ls_j) \wedge (ghq_i = ghq_j) \Leftrightarrow r_i = r_j$.

Notice that as long as the bilateral ranking between individuals in the sample is preserved, any strict monotonic transformation of the constructed variable would represent the data equally well. This implies that the constructed variable has an ordinal nature with the potential of 259 ordered levels[34].

To examine the suitability of the two self-reported well-being measures for the construction of the composite measure, the two are jointly modelled in subsequent subsections through the use of a bivariate ordinal regression model. In addition, the aforementioned methodology is used to create an alternative well-being variable which acts as a robustness check to the measure described in the current subsection.

5.3 Bivariate ordinal regression model

The bivariate ordinal regression model provides a framework for the joint modelling of two ordinal variables under the assumption that a latent variable with an unobserved continuous distribution underlies each ordinal variable. In this paper's application, self-reported well-being constitutes the ordinal variable of interest, with the true, unobserved value of well-being representing the latent variable. The two unobserved univariate distributions constitute the marginal distributions of the unobserved continuous bivariate distribution used to capture the dependence structure between the two latent variables.

For individual i, the latent variable y_{ji}^* for $j \in \{1,2\}$ is determined such that:

[33] Recall the transformation performed on the GHQ measure such that a higher GHQ score represents a higher level of well-being.
[34] The Cartesian product of $\{1, \dots, 7\} \times \{0, \dots, 36\}$.

$$y_{ji}^{*} = \boldsymbol{x}_{ji}'\boldsymbol{\beta}_j + \varepsilon_{ji}. \quad (1)$$

It is therefore implied that each of the latent variables depends on a vector of explanatory variables $\boldsymbol{x}_{ji}$ in a linear fashion through a conformable vector of parameters $\boldsymbol{\beta}_j$, and an additive stochastic component ε_{ji}.

The observable counterpart y_{ji} is determined such that:

$$y_{ji} = r \Leftrightarrow \gamma_{rj} \leq y_{ji}^{*} < \gamma_{r+1j},$$

where $r \in \{1, \dots, R_j\}$.

γ_{rj} and γ_{r+1j} represent threshold values in the underlying variable y_{ji}^{*} for category r. In order to cover the entire support of y_{ji}^{*}, γ_{1j} represents $-\infty$ and γ_{R_j+1j} represents $+\infty$. The observable counterpart y_{ji} has $|\{1, \dots, R_j\}|$ levels.

It is often the case that the joint distribution between the two stochastic components is assumed to be a bivariate normal distribution with a dependence parameter represented by the correlation coefficient. In that case the model boils down to the bivariate ordered probit model.

However, in this case a more general version of the model is used, as provided by Hernández-Alava and Pudney (2016), which nests the ordered probit specification. The generality, and thus flexibility, of the model presented by the authors comes in the form of the specification of the dependence structure between the two stochastic components ε_{ji}. By exploiting Sklar's theorem, the authors express the bivariate distribution function $F(\varepsilon_{1i}, \varepsilon_{2i})$ such that:

$$F(\varepsilon_{1i}, \varepsilon_{2i}) = C\left(F_{\varepsilon_{1i}}(\varepsilon_{1i}), F_{\varepsilon_{2i}}(\varepsilon_{2i})\right).$$

C represents the unique copula function which captures the dependence structure between the continuous marginal distribution functions $F_{\varepsilon_{ji}}(\varepsilon_{ji})$. In this context, the choice of a Gaussian copula in combination with standard normal marginal distribution functions is equivalent to the standard bivariate ordered probit model.

The flexibility in the specification choice is present both in the choice of the copula, as well as in the choice of the marginal distributions for each stochastic component. The choice between the Independence, Clayton, Frank, Gumbel, and Joe copulas can be made. Each copula represents a distinct shape for the bivariate distribution of two variables, which moves away

from the commonly used normal distribution. In addition, marginal distributions can be modelled as normal mixtures.

A normal mixture distribution in this case is defined as the linear combination of two normal distributions. More formally, for $F_{\varepsilon_{ji}}(\varepsilon_{ji})$:

$$F_{\varepsilon_{ji}}(\varepsilon_{ji}) = \delta_j \Phi\left(\frac{\varepsilon_{ji}-\mu_{j1}}{\sigma_{j1}}\right) + (1-\delta_j)\Phi\left(\frac{\varepsilon_{ji}-\mu_{j2}}{\sigma_{j2}}\right).$$

Φ represents the standard normal cumulative distribution function. δ_j is the mixing probability. μ_{j1} and μ_{j2} are location parameters, and σ_{j1} and σ_{j2} are scale parameters which satisfy the conditions:

$$\delta_j\mu_{j1} + (1-\delta_j)\mu_{j2} = 0, \text{ and } \delta_j(\sigma_{j1}^2 + \mu_{j1}^2) + (1-\delta_j)(\sigma_{j2}^2 + \mu_{j2}^2) = 1.$$

Flexibility in the marginal distributions of the stochastic components permits the modelling of features such as skewness and bimodality, when compared to the case of using the normal distribution. Flexibility in the distributions used to capture the dependence between the stochastic components permits the modelling of features such as (asymmetric) tail dependence, when compared to the case of using the normal distribution[35].

5.4 Life satisfaction and GHQ bivariate ordinal regression model

Life satisfaction and GHQ are jointly studied through the bivariate ordinal regression model to examine the extent to which the two self-reported measures represent similar underlying latent variables. Loosely speaking, an attempt is made to assess the extent to which the two survey questions elicit similar information from the individuals. This is done on the basis of the estimated associations between a set of explanatory variables and each one of the latent variables. Therefore, it is the case that $\boldsymbol{x}_{1i} = \boldsymbol{x}_{2i}$ for all $i \in \{1, \dots, n\}$ in specification (1) provided in subsection 5.3 where it is assumed that life satisfaction is indexed as 1, and GHQ is indexed as 2.

In addition, the estimated model is used to derive the conditional distribution of life satisfaction given the value of GHQ and the set of explanatory variables. This is used to construct a ranking of well-being which is almost entirely tie-free. The comparison between this ranking and the composite measure of subsection 5.2 is used as evidence in favour of the validity of the

[35] Estimation of the bivariate ordinal regression model is performed through the use of the bicop command offered by Hernández-Alava and Pudney (2016) for the statistical software Stata.

composite measure in capturing well-being while still providing a way of breaking the ties in self-reported life satisfaction.

The bivariate ordinal regression model is fitted using maximum likelihood estimation[36]. Before examining the estimates of the coefficient vectors $\boldsymbol{\beta}_j$ associated with the vector of explanatory variables $\boldsymbol{x}_{ji}$, the structure of dependence between the stochastic components ε_{ji} is presented, as characterised by the identified copula function and the associated estimated parameters.

The Frank copula is identified as the copula function of best fit for the data. The Frank copula cumulative distribution function (CDF) C^F is given by:

$$C^F(\varepsilon_1, \varepsilon_2; \theta) = -\left(\frac{1}{\theta}\right) \ln\left\{1 + \frac{(e^{-\theta\varepsilon_1}-1)(e^{-\theta\varepsilon_2}-1)}{e^{-\theta}-1}\right\} \text{ for } \theta \neq 0 \text{ and } (\varepsilon_1, \varepsilon_2) \in [0,1]^2,$$

where $C^F(\varepsilon_1, \varepsilon_2; \theta) = \varepsilon_1\varepsilon_2$ for $\theta \to 0^+$ and $(\varepsilon_1, \varepsilon_2) \in [0,1]^2$.

$C(\varepsilon_1, \varepsilon_2) = \varepsilon_1\varepsilon_2$ for $(\varepsilon_1, \varepsilon_2) \in [0,1]^2$ is the Independence copula.

The estimate of θ in this case is approximately 2.677. This is a value which corresponds to a theoretical Kendall's tau value of approximately 0.278, indicating positive association. For the purpose of illustration, *Figure 1* provides a visual representation of the Frank copula theoretical CDF and probability density function (PDF) for the case of $\theta = 2.677$.

Figure 1: Frank copula theoretical CDF (left) and PDF (right) for $\theta = 2.677$.

[36] Likelihood-based information criteria AIC and BIC are used to choose between different specifications sequentially. Firstly, the choice of the copula distribution function which provides the best fit for the data is made based on a model which assumes that each of the stochastic components, ε_{ji}, has a standard normal distribution. Frank copula is preferred based on both AIC and BIC. Using the copula of best fit, the choice between the aforementioned specification, and the specification which assumes a common normal mixture distribution for the stochastic components is made. The latter is preferred based on both AIC and BIC. Finally, a choice is made between the common normal mixture specification, and the specification which assumes a different normal mixture distribution for each stochastic component. The latter one is preferred based on both AIC and BIC.

In relation to the Gaussian copula, for the Frank copula dependence is strongest in the middle of the distribution and weaker in the tails. This implies that an appropriate measure of association (e.g. correlation) would be higher in magnitude when considering values closer to the means of the two variables, as opposed to extreme values of the two variables.

In addition to the copula identification and estimation, *Table 2* presents the estimates of the parameters associated with the normal mixture distributions of the two stochastic components, following the notation of subsection 5.3.

Based on the estimated parameters, the theoretical density functions for the two normal mixture distributions of the stochastic components are presented in *Figures 2* and *3*. *Figure 2* accounts for the stochastic component of life satisfaction, and *Figure 3* for the stochastic component of GHQ. In each of the two illustrations, the standard normal density is also included as a point of reference.

Figures 2 and *3* act as indications for the difference of each distribution from the standard normal distribution. These differences are taken as support for the methodology used. The standard normal distribution was a possible choice in the first place. The fact that both estimated distributions differ from the standard normal speaks to the flexibility of the methodology.

Table 2: Parameter estimates for normal mixture distributions of stochastic components.

Parameter	**Estimate**	**Standard error**
δ_1	0.804	0.022
μ_{11}	0.068	0.024
μ_{12}	-0.281	0.110
σ_{11}^2	0.385	0.039
σ_{12}^2	3.431	0.303
δ_2	0.865	0.039
μ_{21}	-0.135	0.017
μ_{22}	0.860	0.324
σ_{21}^2	0.622	0.039
σ_{22}^2	2.555	0.325

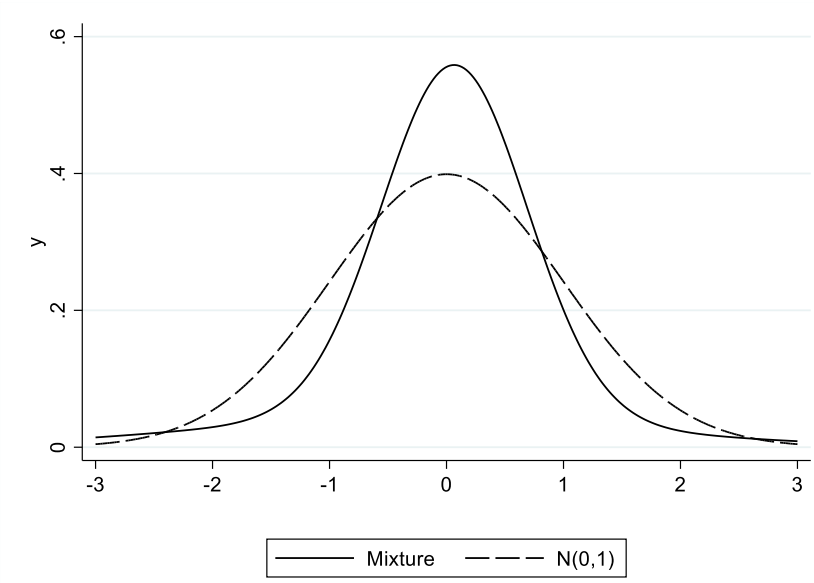

Figure 2: Life satisfaction stochastic component probability density (y) function represented by Mixture.

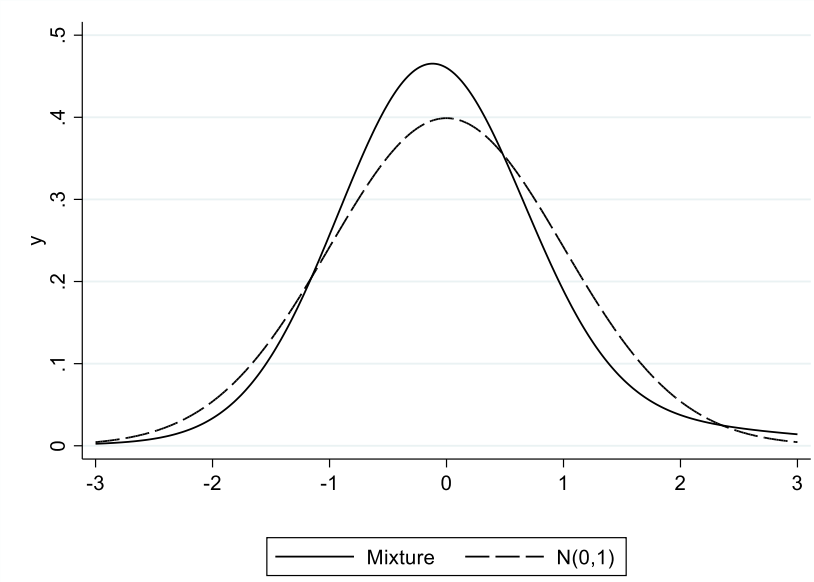

Figure 3: GHQ stochastic component probability density (y) function represented by Mixture.

5.5 Similarities between life satisfaction and GHQ

The variables incorporated in the analysis as explanatory variables in each of the vectors $\boldsymbol{x}_{ji}$ can be divided into two categories. Those that measure socio-economic characteristics, and those that measure personality traits.

Socio-economic characteristics include age, the natural logarithm of equivalised household income[37], and number of own children in household, as well as sets of dummies for economic activity, country of residence, gender, marital status, highest educational qualification, general health[38], and ethnicity. In addition, year dummies are included based on the calendar year during which the interview is carried out.

The other set of covariates included in $\boldsymbol{x}_{ji}$ are the variables aimed at capturing the personality characteristics of the individuals interviewed. The variables available in Understanding Society represent the Big Five personality traits. Studies such as that of Goldberg (1990), and McCrae and John (1992) are supportive of the general applicability of this method of capturing the overall structure of an individual's personality. The five dimensions describing an individual's personality include extraversion (e.g. being outgoing and talkative), agreeableness (e.g. being trusting and kind), conscientiousness (e.g. being responsible and thorough), neuroticism (e.g. being anxious and worrying), and openness (e.g. being creative and curious). Personality characteristics are captured in wave 3 of the Understanding Society survey. The variables representing each of the five dimensions are recorded on a 7-point Likert scale. For the subsequent analysis, these variables are assumed to be ordinal in nature, and are thus included as a set of dummies in the specification.

Table 3 provides the coefficient estimates for $\boldsymbol{\beta}_1$, the coefficient vector linking the explanatory variables to the latent variable underlying life satisfaction, and the coefficient estimates for $\boldsymbol{\beta}_2$, the coefficient vector linking the explanatory variables to the latent variable underlying GHQ. Note that the sample size used to estimate the bivariate regression model is smaller than the one used to identify and estimate the vine copula due to excluding observations with missing values in the explanatory variables used. Summary statistics of the variables used in the bivariate ordinal regression model are provided in Appendix C.

A comparison across the two latent variables in terms of the association with the conditioning explanatory variables can provide an indication of the similarity between the two unobserved variables which are supposed to capture the level of well-being. If it is the case that the two ordinal variables provide information on approximately the same unobserved aspect of life,

[37] Pfaff (2013), in a study aiming to explore the features involved in the analysis of life satisfaction when using survey data, promotes the use of equivalised household income as it accounts for household size and composition. Therefore, the square root scale is used (OECD, 2011). In addition, the income variable is adjusted for inflation so that it represents real income. UK inflation data is available by the Office for National Statistics on https://www.ons.gov.uk/economy/inflationandpriceindices.

[38] Health assessment is based on a subjective, self-reported measure.

then, at least qualitatively, the associations implied by the coefficient estimates for $\boldsymbol{\beta}_j$ should not be very different across the two.

Table 3: Bivariate ordinal regression model of life satisfaction and GHQ.

Variable	Life satisfaction	GHQ
Year (Default: 2010)		
2011	-0.0262	0.00401
	(0.0216)	(0.0225)
2012	-0.0586	-0.117
	(0.0750)	(0.0701)
Job Status (Default: Self-employed)		
Paid employment	0.0409	-0.0254
	(0.0385)	(0.0412)
Unemployed	-0.0654	-0.250**
	(0.0736)	(0.0832)
Retired	0.448***	0.176***
	(0.0519)	(0.0489)
On maternity leave	0.420**	-0.0783
	(0.132)	(0.165)
Family care	0.167*	-0.0178
	(0.0655)	(0.0649)
Full-time student	0.487***	0.0795
	(0.0886)	(0.0919)
Long-term sick or Disabled	-0.208*	-0.354***
	(0.0985)	(0.0967)
Government training scheme	0.657**	2.860***
	(0.222)	(0.175)
Unpaid, family business	0.0127	0.0354
	(0.285)	(0.346)
Doing something else	0.116	-0.150
	(0.287)	(0.211)
Health (Default: Excellent)		
Very good	-0.164***	-0.164***
	(0.0329)	(0.0341)
Good	-0.464***	-0.411***
	(0.0380)	(0.0402)
Fair	-0.745***	-0.813***
	(0.0531)	(0.0549)
Poor	-1.265***	-1.263***
	(0.109)	(0.0971)
Country (Default: England)		
Wales	0.0176	0.0112
	(0.0357)	(0.0381)
Marital Status (Default: Single)		
Married	0.134***	-0.0133
	(0.0374)	(0.0426)
Same-sex civil partnership	0.398	-0.263
	(0.226)	(0.198)
Separated	-0.204*	-0.190*
	(0.0905)	(0.0860)
Divorced	-0.163**	-0.162**
	(0.0519)	(0.0559)

Widowed	0.0263	-0.0633
	(0.0585)	(0.0623)
Separated from civil partner	-0.273***	-0.758***
	(0.0702)	(0.0805)
Living as couple	0.0984*	-0.0551
	(0.0430)	(0.0497)
Children number	-0.0376**	-0.0456**
	(0.0145)	(0.0164)
Education (Default: Degree)		
Other higher degree	-0.0202	0.0171
	(0.0334)	(0.0380)
A-level etc	-0.0124	0.00232
	(0.0307)	(0.0349)
GCSE etc	-0.0173	0.0328
	(0.0328)	(0.0338)
Other qualification	0.0236	0.0303
	(0.0431)	(0.0439)
No qualification	0.161***	0.100*
	(0.0460)	(0.0447)
Logarithm of income	0.0946***	0.0315
	(0.0218)	(0.0192)
Agreeableness (Default: 1)		
2	-1.227*	-0.367
	(0.594)	(0.872)
3	-1.241*	-0.343
	(0.556)	(0.853)
4	-1.284*	-0.507
	(0.552)	(0.852)
5	-1.225*	-0.446
	(0.551)	(0.852)
6	-1.194*	-0.447
	(0.551)	(0.852)
7	-1.087*	-0.422
	(0.552)	(0.852)
Conscientiousness (Default: 1)		
2	0.706	-0.522
	(0.588)	(0.563)
3	0.940	-0.169
	(0.554)	(0.550)
4	1.011	-0.203
	(0.550)	(0.549)
5	1.070	-0.164
	(0.549)	(0.548)
6	1.083*	-0.129
	(0.550)	(0.548)
7	1.156*	-0.0889
	(0.550)	(0.548)

Extraversion (Default: 1)		
2	0.410**	0.220
	(0.140)	(0.133)
3	0.365**	0.234
	(0.135)	(0.128)
4	0.473***	0.325*
	(0.135)	(0.127)
5	0.520***	0.333**
	(0.135)	(0.127)
6	0.554***	0.376**
	(0.136)	(0.128)
7	0.611***	0.404**
	(0.140)	(0.133)
Neuroticism (Default: 1)		
2	-0.297***	-0.329***
	(0.0566)	(0.0510)
3	-0.455***	-0.672***
	(0.0574)	(0.0525)
4	-0.589***	-0.958***
	(0.0594)	(0.0571)
5	-0.727***	-1.202***
	(0.0638)	(0.0651)
6	-0.805***	-1.390***
	(0.0740)	(0.0786)
7	-0.946***	-1.560***
	(0.0911)	(0.102)
Openness (Default: 1)		
2	0.131	0.0953
	(0.125)	(0.126)
3	0.0499	0.101
	(0.119)	(0.120)
4	0.0751	0.116
	(0.118)	(0.118)
5	0.0788	0.0947
	(0.119)	(0.119)
6	0.00344	-0.000430
	(0.120)	(0.120)
7	0.0590	0.0191
	(0.129)	(0.129)
Sex (Default: Male)		
Female	0.0753***	-0.110***
	(0.0226)	(0.0247)
Racial group (Default: British)		
Irish	-0.0335	-0.0419
	(0.137)	(0.106)
Gypsy or Irish traveller	-9.105***	-0.220
	(0.611)	(0.254)
Any other white background	-0.0502	-0.00738
	(0.0629)	(0.0698)

White and black Caribbean	-0.209	-0.0575
	(0.233)	(0.229)
White and black African	-0.0767	-0.260
	(0.190)	(0.234)
White and Asian	-0.192	0.0954
	(0.150)	(0.131)
Any other mixed background	0.187	0.363
	(0.209)	(0.378)
Indian	-0.408***	-0.163
	(0.103)	(0.119)
Pakistani	-0.201	-0.156
	(0.178)	(0.146)
Bangladeshi	-0.234	-0.0147
	(0.308)	(0.243)
Chinese	-0.289	0.0656
	(0.268)	(0.285)
Any other Asian background	-0.00450	-0.0281
	(0.159)	(0.181)
Caribbean	-0.328	-0.327
	(0.184)	(0.288)
African	-0.267	-0.265
	(0.192)	(0.211)
Any other black background	-0.693*	-0.420*
	(0.299)	(0.194)
Arab	0.776**	0.282
	(0.268)	(0.202)
Any other ethnic group	0.0411	-0.0676
	(0.321)	(0.406)
Age	0.00103	-0.000538
	(0.00113)	(0.00132)
Observations	7,317	

*Notes: Robust standard errors in parentheses; *p-value < 0.05, **p-value < 0.01, ***p-value < 0.001. Sample size smaller than the one used to identify and estimate the vine copula due to excluding observations with missing values in the explanatory variables used.*

The first thing to note is that in both cases the dummy variables associated to health are highly significant. Based on the coefficient estimates, it is implied that a lower level of health is associated with a lower level of well-being regardless of which measure of well-being is chosen. This finding agrees with studies such as Gerdtham and Johannesson (2001), and Clark and Oswald (2002).

When it comes to job status, a common element in both tables is the case of the retired individuals exhibiting a significant increase in underlying well-being relative to those who are self-employed. In addition, those classified as long-term sick or disabled exhibit a significant decrease in underlying well-being relative to self-employed individuals. The latter could be linked back to the health component of well-being. One important difference is the case of unemployment associated with a significant decrease in underlying well-being when considering GHQ as opposed to life satisfaction where it appears as insignificant. The happiness literature shows consistency in terms of the significant negative impact of unemployment (Ferrer-i-Carbonell, 2013; Clark, 2018). However, it should be noted that estimated coefficients are negative in both cases. Differences also exist in terms of coefficients which appear significant for one of the measures but not the other. For individuals on maternity leave, being occupied with family care, and full-time students the estimated coefficients are positive and significant for life satisfaction, as opposed to the case of GHQ. For those being part of a government training scheme, the estimated coefficient is positive and significant for GHQ, as opposed to the case of life satisfaction. However, given that these groups of individuals represent categories which capture only a small proportion of the sample, individually and on aggregate, the estimated coefficients may not be the best form of evidence in support of the aforementioned inferences.

With regard to the number of own children in household, there is a significant negative association between the number of children and underlying well-being for both GHQ and life satisfaction based on the estimated model. This finding is consistent with the literature in general (Ferrer-i-Carbonell, 2013).

Furthermore, the association between the level of education and well-being appears to be similar across the two measures. Every qualification dummy variable is coupled with an insignificant coefficient suggesting that any level of education below that of a degree is not significantly associated with well-being regardless of the measure considered. However, for both measures there is a significant positive estimate for the coefficient associated with the

dummy variable accounting for individuals with no qualifications. In a previous study which uses the GHQ measure Clark and Oswald (1994) demonstrate a negative association of well-being with the level of education. They speculate that this might be some type of comparison effect generated by high aspirations.

In relation to marital status, it is apparent that for both life satisfaction and GHQ being divorced or separated is associated with a significantly lower level of well-being as opposed to the case of being single. The main difference appears to be in the case of individuals who have a partner (i.e. married, and living as a couple). For life satisfaction, the estimated model suggests a significant positive association between well-being and having a partner as opposed to being single, which appears to be the most common finding in literature (Clark, 2018), whereas for GHQ the coefficients coupled with the relevant dummy variables are insignificant.

The two main differences between the two measures when it comes to variables outside those representing personality characteristics seem to be the estimated coefficients associated with gender, and the natural logarithm of equivalised household income. When considering life satisfaction, the estimated model implies that being female is associated with a significantly higher underlying level of well-being as opposed to being male, whereas when considering GHQ, the estimated model implies the opposite association of gender with well-being. Such inconsistencies which depend on the self-reported measure used are not uncommon in literature (Ferrer-i-Carbonell, 2013; Clark, 2018). In addition, for life satisfaction the natural logarithm of income exhibits a significant positive association with well-being, whereas for GHQ the association can be interpreted as insignificant. The impact of income on well-being is extensively studied in literature without reaching a common consensus (examples include Ferrer-i-Carbonell, 2005; Senik, 2004; and Boyce *et al.*, 2010).

Moving on to the sets of dummy variables representing personality characteristics, there are strong similarities between life satisfaction and GHQ. Firstly, in both cases the dummy variables associated with extraversion are significant[39]. Based on the coefficient estimates, it is implied that a higher level of extraversion is associated with a higher level of well-being regardless of which measure of well-being is chosen out of the two. Furthermore, in both cases the dummy variables associated with neuroticism are highly significant. Based on the coefficient estimates, it is implied that a higher level of neuroticism is associated with a lower level of well-being regardless of which measure of well-being is chosen out of the two. The

[39] Apart from the first two for the case of GHQ.

main difference between the two measures has to do with the sets of coefficients associated with agreeableness and conscientiousness. For agreeableness, the coefficients appear to be marginally significant (individual coefficient significance at the 5% significance level) when considering life satisfaction, whereas they can be interpreted as insignificant in the case of GHQ. For conscientiousness, this is the case for the coefficients associated with level 6 and level 7. The coefficients linked to openness are individually insignificant regardless of the well-being measure under consideration.

Overall, there appears to be a substantial level of similarity between the latent variables underlying the two measures, as captured by the association of the two with a common set of explanatory variables. The apparent differences between the two could be attributed to the particular aspect of well-being captured by each one. Even though many times in literature the general notion of well-being is approximated by both, intuitively it can be argued that life satisfaction is a more inclusive notion that might encompass the aspects captured by GHQ. Despite the similarity acting as evidence in favour of the assumption that the two measures capture approximately the same form of underlying well-being, another check is considered to reinforce the case in favour of the composite measure construction of subsection 5.2. This is the reasoning behind the next subsection. Another well-being measure is constructed that is founded on the bivariate ordinal regression model of the current subsection. As such, the construction in the next subsection does not assume that the two measures capture exactly the same concept. This measure can be used as a robustness check for the original measure of subsection 5.2.

5.6 Alternative well-being measure

In this subsection a composite measure of well-being is constructed which is different from the one in subsection 5.2. This is done by using the estimated bivariate ordinal regression of the previous subsection. The bivariate regression allows the two measures of life satisfaction and GHQ to be modelled jointly by assuming that the latent variables that underlie each measure are not the same. Therefore, it offers more flexibility than the assumption that both measures capture the same concept of well-being. As such, a measure constructed on the basis of the bivariate regression estimation can act as a robustness check for the original composite well-being measure of subsection 5.2.

The estimated bivariate ordinal regression model permits the evaluation of the probability that an observation takes a particular value with respect to life satisfaction, conditional on GHQ

taking a specific value, and conditional on the vector of explanatory variables. This conditional probability is expressed as:

$$Pr(y_{1i} = r|y_{2i} = s, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}) = \frac{Pr(y_{1i}=r, y_{2i}=s|\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i})}{\sum_{k=1}^{7} Pr(y_{1i}=k, y_{2i}=s|\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i})},$$

for all $i \in \{1, \dots, n\}$ where $r \in \{1, \dots, 7\}$ and $s \in \{0, \dots, 36\}$.

Based on this capability of the estimated model, an attempt is made to break the ties that exist in self-reported life satisfaction. This is done by using information provided by the secondary self-reported well-being variable, and the set of explanatory variables, to break the ties in the primary self-reported well-being variable. Using the conditional probability evaluation displayed above, the following conditional probability is generated for each individual i reporting $y_{1i} = r$ and $y_{2i} = s$:

$$p_i = Pr(y_{1i} \geq r|y_{2i} = s, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}) = \sum_{z=r}^{7} Pr(y_{1i} = z|y_{2i} = s, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}),$$

where $r \in \{2, \dots, 7\}$.

In the case that $r = 1$:

$$p_i = Pr(y_{1i} > 1|y_{2i} = s, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}) = \sum_{z=2}^{7} Pr(y_{1i} = z|y_{2i} = s, \boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}).$$

For any two individuals i, j where $i \neq j$ such that $y_{1i} = y_{1j}$, individual i is assumed to have a higher level of well-being if and only if $p_i > p_j$. They are assumed to have the same level of well-being if and only if $p_i = p_j$. Otherwise, individual j is assumed to have a higher level of well-being. Like in the case of the composite measure construction, this approach generates an ordinal ranking of individuals' well-being, where in this case information based on GHQ and explanatory variables is used through the bivariate ordinal regression model. It is worth noting that this approach generates an ordinal ranking of well-being with only nine cases of two-way ties between individuals in a sample of 7,317 individuals used to estimate the model.

Generating the aforementioned, almost entirely tie-free, ordinal ranking of individuals' well-being is equivalent to generating a new well-being variable z_i such that $z_i = y_{1i} + p_i$ and assume that z_i has an ordinal nature. The generated variable can be used to examine the validity of the constructed composite well-being measure in subsection 5.2. If the original composite variable r_i capturing the well-being ranking is shown to be 'close' to the generated variable z_i, then, on the basis of the specified bivariate well-being model, there is evidence in favour of the informational validity of the composite measure in subsection 5.2.

The two variables considered in this case have an ordinal nature such that only the ranking of individuals matters with respect to each individual variable, and not the absolute value that is assigned to each individual. As such, to measure how 'close' the two variables are, a measure of rank correlation is chosen. The rank correlation is captured by Kendall's tau.

The theoretical value of Kendall's tau (Kendall's τ) between two continuous random variables x_1 and x_2 is such that:

$$\tau(x_1, x_2) = P\left((x_{1i} - x_{1j})(x_{2i} - x_{2j}) > 0\right) - P\left((x_{1i} - x_{1j})(x_{2i} - x_{2j}) < 0\right),$$

where (x_{1i}, x_{2i}) and (x_{1j}, x_{2j}) in this particular exposition are independent random vectors with a common distribution identical to that of (x_1, x_2).

In order to obtain an empirical version of Kendall's tau, the concepts of concordant, discordant, and extra pairs are needed[40]. For any pair of observations (x_{1i}, x_{2i}) and (x_{1j}, x_{2j}) of a random sample $\{x_{1i}, x_{2i}, i = 1, \dots, n\}$ from the joint distribution of (x_1, x_2), the pair is defined as *concordant* if $x_{1i} > x_{1j}$ and $x_{2i} > x_{2j}$, or $x_{1i} < x_{1j}$ and $x_{2i} < x_{2j}$[41]. The same pair is defined as *discordant* if $x_{1i} < x_{1j}$ and $x_{2i} > x_{2j}$, or $x_{1i} > x_{1j}$ and $x_{2i} < x_{2j}$. Furthermore, if a pair is neither concordant nor discordant, it is called *extra x_1 pair* if $x_{1i} = x_{1j}$, or *extra x_2 pair* if $x_{2i} = x_{2j}$. Note that there are $\binom{n}{2} = \frac{n(n-1)}{2}$ possible pairs of observations in a sample of size n. n_c denotes the number of concordant pairs in a sample of size n. n_d denotes the number of discordant pairs. n_1 and n_2 denote the number of extra x_1 pairs and extra x_2 pairs, respectively. An observation for which both $x_{1i} = x_{1j}$ and $x_{2i} = x_{2j}$ is accounted for in both n_1 and n_2.

The empirical version of Kendall's tau τ_a which does not account for ties is defined such that:

$$\tau_a = \frac{n_c - n_d}{n}.$$

The empirical version of Kendall's tau τ_b which accounts for ties is defined such that:

$$\tau_b = \frac{n_c - n_d}{\sqrt{n - n_1}\sqrt{n - n_2}}.$$

[40] The exposition regarding Kendall's tau is mainly based on the one by Czado (2020).
[41] Note that for the exposition regarding the theoretical Kendall's tau (x_{1i}, x_{2i}) for $i \in \{1, \dots, n\}$ represent random vectors, whereas in the exposition regarding the empirical version of Kendall's tau they represent realisations of the aforementioned random vectors.

Note how the two values coincide in the case that there are no ties in the sample. Otherwise $\tau_a < \tau_b$.

In the case of the two variables considered here, i.e. r_i and z_i, the $\tau_a = 0.880$ and $\tau_b = 0.893$. Both values act as indication that the two generated ordinal rankings for well-being are very 'close' from a rank correlation perspective. As such, it can provide evidence in favour of the validity of the composite measure of subsection 5.2.

6. RESULTS

This section presents the results from the estimation of the regular vine copula. The vine copula attempts to capture the multivariate association between variables including well-being, biomarkers, and physiological measures.

Instead of jointly modelling the association of the entire set of variables all at once, the procedure involves modelling the association between two variables at a time. If there is a significant relationship between the two variables, then the appropriate type of copula is chosen to represent it. Otherwise, the two variables are assumed to be independent. An independence test is used to determine whether the relationship between two variables is significant.

Each type of copula represents a distinct shape for the bivariate distribution of the two variables, which can depart from the commonly used shape of the normal distribution. Therefore, each copula represents a different form of association. The set of copula families considered include the Independence, Gaussian, Student t, Clayton, Gumbel, Frank, and Joe families. Gaussian and Student t represent the well-known homonymous distributions. For demonstration purposes, *Figure 4* presents simulated bivariate samples (1,000 observations each) from different types of copulas, including the Gaussian (normal) copula. For all simulated samples the underlying association is assumed to be such that Kendall's tau is 0.7.

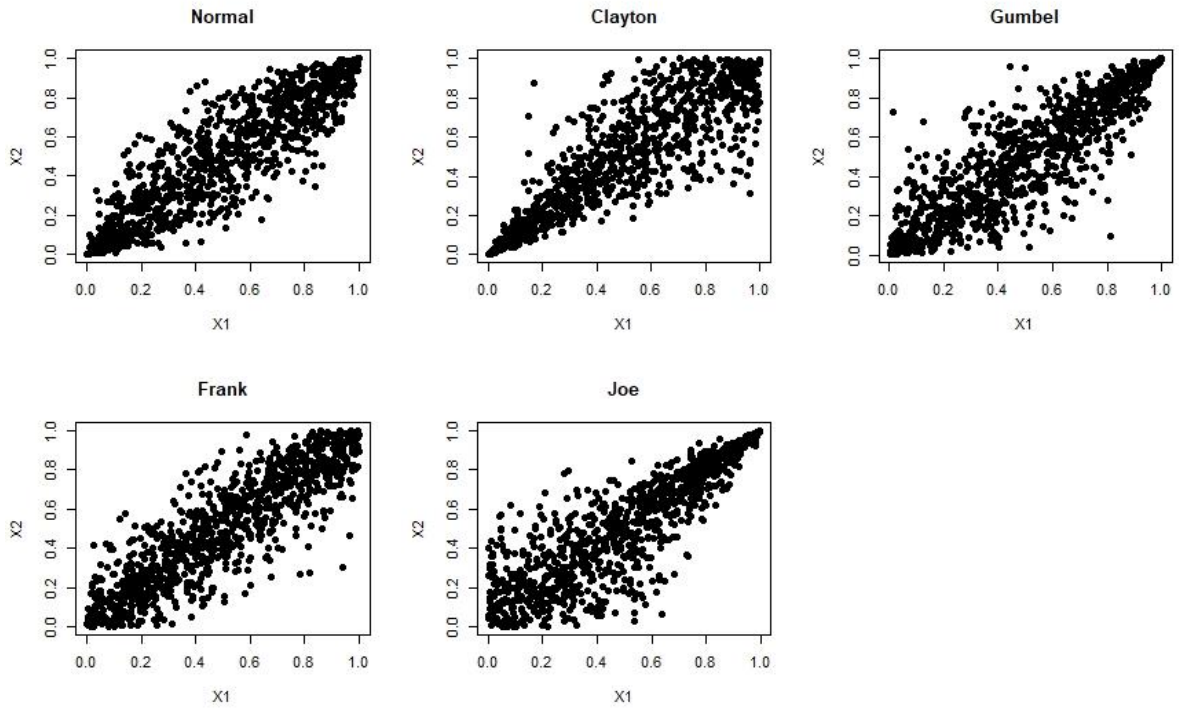

Figure 4: Simulated bivariate samples (1000 observations each) from five different copulas with a Kendall's tau value of 0.7.

Additional copula families which represent combinations or extensions of the aforementioned families are considered. These include the BB1 (Clayton-Gumbel), BB6 (Joe-Gumbel), BB7 (Joe-Clayton), BB8 (Joe-Frank), and Tawn type 1 and 2 families (Gumbel extensions). Most of the copula families considered are designed to capture positive relationships. To allow for more flexibility, rotations of the copulas are also allowed. Rotations can be by 90^o, 180^o, and 270^o. For example, a 90^o rotation of the Clayton copula can be used to capture negative associations.

The relationship between any two variables in our set will either be unconditional or conditional, i.e. given the values of other variables. As an analogy example, $F(X, Y)$ represents the unconditional distribution between any variables X and Y, whereas $F(X, Y|Z)$ represents the conditional distribution between X and Y given the value of Z. For each conditional association, the set of conditioning variables is chosen as part of Dißmann's algorithm which is presented in detail in Appendix B. The algorithm has a sequential nature and prioritizes the modelling of the strongest associations early in the estimation procedure. The strength is judged based on the absolute value of Kendall's tau. For any variable, an association is modelled given the values of all the other variables with which the association is already modelled earlier in the estimation procedure. Therefore, any association modelled in the first step of the algorithm is unconditional.

For ease of exposition, this section only discusses the main implications of the estimated associations between the different variables with well-being. The formal representation of the estimated regular vine copula is provided in Appendix D. In short, a significant unconditional or conditional association between well-being and a biomarker or a physiological measure offers support for the informational validity of the self-reported well-being variable due to the objective way of measuring the rest of the variables.

To avoid using lengthy and technical names for some of the biomarkers and physiological measures, *Table 4* presents some abbreviations which are used in this section.

Table 4: Abbreviations.

Variable (units of measurement)	Abbreviation
Subjective well-being (-)	*drank*
Height (cm)	*height*
Weight (kg)	*weight*
Forced vital capacity (L)	*htfvc*
Albumin (g/L)	*alb*
Dehydroepiandrosterone sulphate ($\mu mlol/L$)	*dheas*
Glycated haemoglobin ($mmol/mol$)	*hba1c*
High-density lipoprotein cholesterol ($mmlol/L$)	*hdl*
C-reactive protein (mg/L)	*hscrp*
Age ($years$)	*age*
Systolic blood pressure ($mmHg$)	*sys*
Diastolic blood pressure ($mmHg$)	*dias*

6.1 Estimated regular vine copula

Table 5 presents the part of the estimated copula which incorporates subjective well-being. The estimated associations between other variables can be found in Appendix D.

Table 5: Subjective well-being associations in the estimated regular vine copula.

Well-being association	Copula	Parameter 1	Parameter 2
drank, age	BB8	1.32	0.87
drank, hba1c*; age*	Clayton 90^o	-0.09	-
drank, sys*; age, hba1c*	Independence	-	-
drank, dias*; age, hba1c, sys*	Gaussian	-0.06	-
drank, dheas*; age, hba1c, sys, dias*	Clayton	0.05	-
drank, htfvc*; age, hba1c, sys, dias, dheas*	Clayton	0.07	-
drank, alb*; age, hba1c, sys, dias, dheas, htfvc*	BB8 180^o	1.23	0.66
drank, hscrp*; age, hba1c, sys, dias, dheas, htfvc, alb*	Independence	-	-
drank, height*; age, hba1c, sys, dias, dheas, htfvc, alb, hscrp*	Independence	-	-
drank, weight*; age, hba1c, sys, dias, dheas, htfvc, alb, hscrp, height*	Independence	-	-
drank, hdl*; age, hba1c, sys, dias, dheas, htfvc, alb, hscrp, height, weight*	Clayton	0.03	-

*Notes: The semicolon separates the variables (in **bold** on the left of the semicolon) for which the conditional association is modelled from the conditioning variables (right of the semicolon).*

The estimated regular vine copula provides information for the association between subjective well-being and each of the biomarkers used in this paper. If the results of this study are to be

used as evidence for the validity of self-reported measures in terms of capturing the level of well-being for an individual, the results should make sense from a medical standpoint, i.e. the direction of any association should indicate that better self-reported well-being is associated with better biological well-being.

Based on the estimated copula, the association between *drank* and each of *sys*, *hscrp*, *height*, and *weight* is modelled by the Independence copula. As such, it can be inferred that there is not a significant relationship between subjective well-being and each of the biomarkers for systolic blood pressure, c-reactive protein, height, and weight. It is worth noting that if the association of each of these variables with well-being was examined in isolation, and not by using the estimated vine copula (i.e. by modelling the unconditional association of each variable with well-being), independence would be rejected.

If the association of each of the aforementioned variables is studied individually, without considering the effect of the rest of the variables through the vine copula, the unconditional associations with *hscrp* and *weight* are negative and significant at the 1% significance level, the one with *height* is positive and significant at the 5% significance level, and the one with *sys* is positive and significant at the 1% significance level. The finding regarding *hscrp* is in agreement with the literature (Hamer and Chida, 2011). A higher quantity of *hscrp* in the blood is associated with a response of the body to inflammation. The findings with regard to height and weight may be considered as intuitively reasonable at first sight. A taller and leaner person experiences a higher level of mental health. The only result which seems to be in contradiction to the literature (Szabo *et al*., 2020) is the one between *drank* and *sys* as it appears to be positive and significant at the 1% significance level. Increased blood pressure is associated with higher risk of cardiovascular disease. However, based on the findings of the vine copula modelling these variables appear to transmit the influence of other biomarkers on well-being when studied on their own.

As far as the rest of the variables are concerned, the estimation suggests a significant **positive** relationship between *drank* and *age*. Based on the two estimated parameters for the BB8 copula, the Kendall's tau which corresponds to the association between the two variables is given by 0.09. There is a difference between the positive association implied by the BB8 copula and the positive association implied by the commonly used normal distribution. Based on

simulated data from both the BB8 copula and the Gaussian copula[42], an ordinary least squares (OLS) estimation of a quadratic model which takes the variable representing well-being as the dependent variable and the one representing age as the explanatory variable suggests that the coefficients on the first order and second order terms are both positive. However, only in the case of the simulated data from the BB8 copula the second order term is significantly positive at the 1% significance level. Therefore, the modelled unconditional distribution of well-being and age based on the estimated copula can support a convex relationship between the two. This result lies close to the literature findings of a U-shaped association of well-being with age. Blanchflower and Oswald (2008b), Glenn (2009), and Frijters and Beatton (2012) are some of the studies arguing whether well-being exhibits convexity across the lifespan of individuals. Similar exercises can be carried out to understand the implications of data originating from different copula families for the conclusions drawn using common methods such as OLS.

The estimation also suggests a significant **negative** relationship between *drank* and *hba1c* given *age*. Based on the estimated parameter, the Kendall's tau which corresponds to the association between the two variables is given by -0.04. The result appears to make sense from a medical standpoint as glycated haemoglobin is a measure of the sugar level in the blood. Higher values can be used in diagnosing diabetes. The finding is also in agreement with the literature in terms of the negative association between subjective well-being and glycated haemoglobin (Tsenkova *et al*., 2000; Poole *et al*., 2019). It is worth noting that if the association between the two variables is examined on its own, and not through the vine copula, it is positive but insignificant even at the 10% significance level. This is based on the same independence test used in vine copula construction to determine whether a bivariate relationship is significant[43].

Based on the estimated regular vine copula, there is a significant **negative** association between *drank* and *dias* given *sys*, *hba1c*, and *age*. The Kendall's tau which corresponds to the conditional association between the two variables is given by -0.04. In this case, the independence test of the unconditional association between the two also provides evidence of a significant negative relationship at the 1% significance level. High blood pressure is linked

[42] 10,000 observations are simulated for each case. The Gaussian copula is assumed to have a parameter corresponding to a Kendall's tau value of 0.09.

[43] For comparison purposes, regressing *drank* on *hba1c*, controlling for *age*, gives an OLS coefficient of -0.013 which is significant at the 1% level. The dependent variable in this estimation is assumed to be on the interval scale (1,8) and the *hba1c* independent variable has a mean of 36.879 with standard deviation of 7.616 as shown in *Table 1*. This analogous OLS approach is reported for all significant relationships identified by the estimated regular vine copula.

to a high chance of cardiovascular disease. Based on the findings of this paper it appears to be the case that the association between diastolic blood pressure and subjective well-being is stronger than that between systolic blood pressure and well-being[44].

The relationship between *drank* and *dheas* given *dias*, *sys*, *hba1c*, and *age* is modelled as well. This is a significant **positive** conditional relationship between well-being and dehydroepiandrosterone sulphate. The Kendall's tau which corresponds to the association between the two variables is given by 0.02. In this case the independence test suggests a significant, but negative, unconditional relationship between the two variables at the 1% significance level. The positive conditional association is in line with the expectation based on what the biomarker accounts for. Dehydroepiandrosterone sulphate is a biomarker for which higher levels are associated with better health. This finding agrees with studies such as Wong *et al*. (2011), and Valtysdottir *et al*. (2003)[45].

Another significant relationship is that between *drank* and *htfvc* given *dheas*, *dias*, *sys*, *hba1c*, and *age*. Based on the estimated parameter there is a **positive** association between the two variables. The relevant Kendall's tau is 0.04. The independence test of the unconditional association in this case suggests an insignificant negative relationship at the 10% significance level. As far as well-being is concerned, the positive conditional association is a reasonable inference as higher values for *htfvc* are linked to better respiratory functioning. This finding agrees with the study of Goracci *et al*. (2008), even though in this case the result applies to a more general group of individuals rather than only those with sarcoidosis[46].

There is a significant conditional **positive** association between *drank* and *alb* given *htfvc*, *dheas*, *dias*, *sys*, *hba1c*, and *age*. Kendall's tau is estimated to be 0.03. Again, the independence test for the unconditional relationship between well-being and albumin suggests an insignificant association between the two at the 10% significance level. This is a finding which also makes sense from a medical point of view as low levels of albumin are linked to possible loss of liver function. The literature also seems to agree with the finding (Schenk *et al*., 2018; Prinsloo *et al*., 2015)[47].

[44] Regressing *drank* on *dias*, controlling for *age*, *hba1c* and *sys*, gives an OLS coefficient of -0.014 which is significant at the 1% level.
[45] Regressing *drank* on *dheas*, controlling for *age*, *hba1c*, *sys* and *dias*, gives an OLS coefficient of 0.008 which is not significant even at the 10% level.
[46] Regressing *drank* on *htfvc*, controlling for *age*, *hba1c*, *sys*, *dias* and *dheas*, gives an OLS coefficient of 0.060 which is significant at the 1% level.
[47] Regressing *drank* on *alb*, controlling for *age*, *hba1c*, *sys*, *dias*, *dheas* and *htfvc*, gives an OLS coefficient of 0.026 which is significant at the 1% level.

The last estimated relationship suggests a significant **positive** conditional association between *drank* and *hdl* given *weight*, *height*, *hscrp*, *alb*, *htfvc*, *dheas*, *dias*, *sys*, *hba1c*, and *age*. The estimated Kendall's tau is 0.02. This agrees with the independence test for the unconditional association between high-density lipoprotein cholesterol and well-being at the 1% significance level. Since *hdl* is considered as the 'good cholesterol', this finding is also supportive of the informational content for self-reported well-being. The finding is consistent with the study by Radler *et al*. (2018)[48].

In general, one of the main points in the current subsection is that no finding contradicts the usefulness of subjective well-being from a medical standpoint. The self-reported measure is capable of capturing information with regard to the underlying biological well-being of individuals as far as the directions of association with the basic biomarkers used in this paper are concerned. For a self-reported measure to capture associations in the 'expected' direction to such an extend provides additional validity to subjective well-being, at least in the form used in the current study.

6.2 Robustness check

One of the main conditions for uniqueness of the copula function, and thus uniqueness of the characterisation of the association between the variables, is that all the variables are continuous. This is equivalent to ranking individuals with respect to any variable, and this ranking being unique for each individual. Unavoidably, due to issues like measurement error or rounding, some of the observations will be tied with each other with respect to one or more variables (Hofert *et al.*, 2018). This issue is inherent for the case of subjective well-being measures since we are assuming that their true nature is continuous, and yet they are recorded on discrete scales. It is then a question of how severe this issue can be when it comes to inferring the appropriate copula function for each pair of variables.

The method based on which ties are dealt in the current paper is such that observations which are tied with respect to the value of a variable are assigned the average rank between them. However, it can be the case that two or more observations are falsely assigned to the same rank value due to loss of information in the manner in which a variable is elicited. For example, two people have a weight of 85kg and we rank them on the same level. However, if we specify that one is 85.1kg and the other is 85.4kg, then we can strictly rank them. For the scope of this

[48] Regressing *drank* on *hdl*, controlling for the rest of the variables, gives an OLS coefficient of 0.113 which is significant at the 1% level.

paper, this issue embodies much of the criticism faced for the use of subjective measures reported on discrete ordinal scales in the analysis of inherently unobserved variables which are assumed to be continuous such as well-being.

Given the risk which underlies inference based on an estimated model using data which contains a significant number of ties for individual variables, it is a question of how to guard, even partially, against such pitfalls. The approach followed in this paper incorporates a method of breaking the ties in the analysis of discrete variables proposed by Denuit and Lambert (2005). This method is also known as jittering and involves adding a continuous random variable to the existing values of the variable for which there are tied observations. The value of the continuous variable added must be sufficiently small such that it does not affect the relative position of observations which are originally untied.

The main idea behind the robustness check described is that ties are broken in a random manner several times, starting each time from the original data. The model selection and estimation steps are carried out in each repetition. The hope is that the inference in terms of the estimated regular vine copula model is not very different each time, with special interest in the bivariate (conditional) associations which incorporate the well-being variable. This process is carried out for 10 repetitions[49].

In addition to the aforementioned procedure, another robustness check is carried out by repeating the model selection and estimation steps using the measure constructed in subsection 5.6. Despite the fact that this measure is generated in the attempt to validate the original measure used in the analysis, it is also used to check for the robustness of the estimated model since it contains a significantly lower number of ties than the original measure.

The estimated regular vine copulas which correspond to the aforementioned robustness checks are presented in Appendix D. It can be seen both in the application of random ranking, and in the use of the alternative well-being measure that the inferences in terms of the associations which consider well-being do not vary much. Minor discrepancies are described in Appendix D.

[49]As in every programming language, a pseudo-random number generator which depends on a number seed is used in the statistical software R to break the ties. This process is carried out for 10 repetitions, each based on a different number seed. The list of number seeds used is 111, 222, 333, 444, 555, 666, 777, 888, 999, and 101010.

6.3 Variations based on gender

Perhaps the main disadvantage of using vine copulas in the manner applied in the current paper is that categorical variables cannot be directly incorporated in the analysis. This is not necessarily an issue in the present study as all of the biomarkers included are inherently recorded on continuous scales. Categorical variables are usually used to capture the socio-demographic characteristics of individuals, such as the marital status or the employment status. These are variables which are traditionally used as determinants of well-being. Still, this does not impose a problem for this study as the primary aim does not involve examining the determinants of well-being. The aim is rather to explore the informational content of subjective well-being with respect to various biomarkers. The bilateral association of well-being with each biomarker is modelled in an environment in which the impact of the rest of the biomarkers is controlled for if deemed necessary. It is then a question of which categorical variables, if any, can affect the inferences made for these bilateral associations. Putting it in other words, it is a question of whether the profile of an individual can influence the informational content of subjective well-being. One example of a variable which accounts for the profile of an individual is *age*. However, since *age* can be assumed to be a continuous variable, it is already controlled for in the estimated vine copula. Another variable which seems a reasonable consideration is the case of the dummy variable for gender, denoted by *sex* from now on. As such, the regular vine copula is estimated twice, once using the subsample consisting of only males, and once using only females. The results are presented in Appendix E. The main interest in this subsection is how the estimated bivariate associations which involve well-being change.

There are some notable differences between the estimated regular vines which use the subsamples based on gender and the original estimation. However, the informational content of subjective well-being is not refuted. Even if some tests described further on provide evidence against the original estimation with respect to the fit for the data, many of the inferences made with respect to the association of well-being with different biomarkers are still there. The positive associations with *htfvc* and *age* are present in all estimations. The negative association with *hba1c* is also present in every estimated model. Furthermore, the negative association with *dias* is present in both models which represent males. Moreover, the positive associations with *alb* and *hdl* are present in both models which represent females.

The first thing to consider when looking at the estimation which uses only male individuals[50] is that the inferences for the associations between *drank* and each one of *age*, *htfvc*, *hba1c*, *hscrp*, *height*, *weight*, and *dias* remain unchanged[51]. One interesting change is the relationship between *drank* and *sys* which is modelled as positive in this case, as opposed to the two being independent originally. Furthermore, independence is also inferred for the relationships between *drank* and each one of *hdl*, *alb*, and *dheas* in this scenario. However, it is not clear as to whether this is because of considering males only, or due to reducing the sample size per se. The selected parametric copula families for some of the associations change as well, but families with similar characteristics are chosen in their place.

By comparing the two estimated models, it is not straightforward to judge the 'distance' between them. As such, a more robust method of determining how much the models differ is required. One solution to this is the case of the Vuong test. The null hypothesis of the likelihood ratio test by Vuong (1989) suggests that the two competing regular vine copulas provide equivalent fits for the given data set of male individuals. The alternative hypothesis suggests that one of the two models is better. If the null hypothesis is rejected, the decision on which model provides the better fit depends on the sign in front of the test statistic. With a p-value of less than 0.001 there is strong evidence in favour of the estimated model presented in Appendix E over the one in subsection 6.1[52]. However, the Vuong test result should be treated with care as the two competing models are not strictly non-nested, but rather overlapping (Czado, 2020). As a result, the non-parametric test by Clarke (2007) is used as well. The Clarke test has essentially the same null hypothesis and follows the same logic as the Vuong test. With a p-value of less than 0.001 there is again strong evidence in favour of the estimated model presented in Appendix E[53].

It should be highlighted again that the fit of the model for the data is not in itself a determinant of whether subjective well-being provides useful information regarding the biological well-being of individuals. The significance of the associations of well-being with the different biomarkers and physiological measures is used for examining the usefulness of well-being, and

[50] The sample size of males is 3,690.
[51] The estimated direction of association is a term used to capture the inference of independence between two variables as well when appropriate.
[52] This is true even in the case of using the Akaike and Schwarz corrections suggested by Czado (2020) for the differing number of parameters between the two models.
[53] Again, the inference made from the test is not affected by the use of the Akaike and Schwarz corrections.

this is just one part of the estimated vine copula. The application of the different tests is merely an exercise for examining how the estimated models differ from an overall perspective.

The same procedure is then carried out using the subsample consisting of only female individuals[54]. The directions of association between *drank* and each one of *age*, *htfvc*, *hba1c*, *hscrp*, *alb*, *height*, *hdl*, and *sys* remain unchanged. The first main change comes in the form of the association between *drank* and *weight*, where the estimated model suggests a negative conditional relationship as opposed to the independence of the original estimation. Furthermore, independence is inferred for the relationships between *drank* and *dheas*, and the one between *drank* and *dias*, whereas in the original estimation there was evidence of positive and negative relationships, respectively. Just like in the case for males, it is not clear whether the sample size reduction is responsible for these last changes. With a p-value of less than 0.001 for both the Vuong and Clarke tests, there is strong evidence in favour of the estimated model in Appendix E over the model presented in subsection 6.1[55].

[54] This accounts for 4,464 observations.
[55] This is true for Akaike and Schwarz corrections as well.

7. CONCLUSION

Subjective, self-reported measures may be used to elicit information about latent variables such as life satisfaction and mental health. They offer a practical approach to incorporate unobserved concepts in quantitative analysis. However, the support for such measures is not unanimous which provides the motivation for this paper. Using data from the UK's Understanding Society survey, the validity of the subjective measures are tested by examining the associations between a composite measure of well-being and a set of biomarkers used to capture the overall state of health (or well-being) for each individual. Each biomarker captures a different aspect of an individual's state of health. If subjective well-being provides 'reasonable' information from a medical standpoint, then the inferred associations should suggest that a higher level of self-reported well-being is associated to a better health state.

A composite measure of well-being is constructed by using an ordinal life satisfaction measure with 7 categories and the GHQ measure of well-being with 37 categories as building blocks. The life satisfaction measure is used as the primary indicator of subjective well-being, and the GHQ measure is used to break the ties (as much as possible) between individuals reporting the same level of life satisfaction. The construction of the composite measure is based on the assumption that the two survey questions elicit similar information from the individuals. This is evaluated in subsection 5.4 on the basis of the estimated associations between a set of explanatory variables and each one of the ordinal variables indicating that the two measures are fairly similar in terms of how they relate to several aspects of an individual's life.

To jointly model the set of biomarkers along with the well-being measure, the regular vine copula is used. Well-being is recorded on a discrete scale but is actually assumed to be a latent continuous variable. In a similar manner to the concept of utility in economics, the choices that individuals make as responses on the discrete scale used to record well-being can remain unchanged for any strict monotonic transformation of the actual unobserved well-being. Copulas allow the examination of the dependence between variables as a separate entity from the marginal distribution of each variable. This means that the characterisation of the associations between well-being with the other variables through copulas will remain unchanged for any strict monotonic transformation of unobserved well-being.

The estimated model suggests that there is evidence in favour of the ability of subjective well-being measures to capture the underlying levels of well-being, at least from a medical standpoint. As far as the biomarkers for glycated haemoglobin, diastolic blood pressure,

dehydroepiandrosterone sulphate, forced vital capacity, albumin, and high-density lipoprotein cholesterol are concerned, the estimated associations are as ‘expected’ in terms of the direction of association. In addition, the inference of conditional independence is made for the association between the composite well-being measure, and each of systolic blood pressure, c-reactive protein, height, and weight. These findings corroborate the usefulness of subjective measures.

Overall, the evidence provided supports the use of subjective well-being measures in literature as they appear to capture useful information with regard to the underlying well-being of individuals. There is room for further research with regard to the approach in the current study. Firstly, the investigation of the same questions as the current paper with the use of a data set with a panel nature would be useful. This is not possible through the Understanding Society survey at the moment as the biomarker data is recorded only once. Furthermore, the examination of the association of such biomarker measures with other self-reported measures which aim to capture alternative aspects of well-being in a multivariate analysis setting would also be interesting.

8. REFERENCES

Aas, K., Czado, C., Frigessi, A. and Bakken, H., 2009. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2), pp.182-198.

Alesina, A. and Giuliano, P., 2015. Culture and Institutions. *Journal of Economic Literature*, 53(4), pp.898-944.

Alesina, A., Di Tella, R. and MacCulloch, R., 2004. Inequality and happiness: are Europeans and Americans different?. *Journal of Public Economics*, 88(9-10), pp.2009-2042.

Barrett-Connor, E., Khan, K. and Yen, S., 1986. A Prospective Study of Dehydroepiandrosterone Sulfate, Mortality, and Cardiovascular Disease. *New England Journal of Medicine*, 315(24), pp.1519-1524.

Becchetti, L., Castriota, S., Corrado, L. and Ricca, E., 2013. Beyond the Joneses: Inter-country income comparisons and happiness. *The Journal of Socio-Economics*, 45, pp.187-195.

Bedford, T. and Cooke, R., 2001. Probability density decomposition for conditionally dependent random variables modeled by vines. *Annals of Mathematics and Artificial Intelligence*, 32, pp.245-268.

Bedford, T. and Cooke, R., 2002. Vines: A new graphical model for dependent random variables. *The Annals of Statistics*, 30(4), pp.1031-1068.

Benzeval, M., Davillas, A., Kumari, M. and Lynn, P., 2014. *Understanding Society: UK Household Longitudinal Study: Biomarker User Guide and Glossary*. Colchester: University of Essex.

Bertrand, M. and Mullainathan, S., 2001. Do People Mean What They Say? Implications for Subjective Survey Data. *American Economic Review*, 91(2), pp.67-72.

Biomarkers Definitions Working Group, 2001. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clinical Pharmacology & Therapeutics*, 69(3), pp.89-95.

Blanchflower, D. and Oswald, A., 2008a. Hypertension and happiness across nations. *Journal of Health Economics*, 27(2), pp.218-233.

Blanchflower, D. and Oswald, A., 2008b. Is well-being U-shaped over the life cycle?. *Social Science & Medicine*, 66(8), pp.1733-1749.

Blanchflower, D. and Oswald, A., 2019. Unhappiness and Pain in Modern America: A Review Essay, and Further Evidence, on Carol Graham's Happiness for All?. *Journal of Economic Literature*, 57(2), pp.385-402.

Bond, T. and Lang, K., 2019. The Sad Truth about Happiness Scales. *Journal of Political Economy*, 127(4), pp.1629-1640.

Boyce, C., Brown, G. and Moore, S., 2010. Money and Happiness. *Psychological Science*, 21(4), pp.471-475.

Brechmann, E. and Schepsmeier, U., 2013. Modeling Dependence with C- and D-Vine Copulas: TheRPackageCDVine. *Journal of Statistical Software*, 52(3).

Brown, S., Gray, D. and Roberts, J., 2015. The relative income hypothesis: A comparison of methods. *Economics Letters*, 130, pp.47-50.

Clark, A. and Oswald, A., 1994. Unhappiness and Unemployment. *The Economic Journal*, 104(424), pp.648-659.

Clark, A., 2015. SWB as a Measure of Individual Well-Being. *PSE Working Papers*.

Clark, A., 2018. Four Decades of the Economics of Happiness: Where Next?. *Review of Income and Wealth*, 64(2), pp.245-269.

Clark, A. and Oswald, A., 2002. A simple statistical method for measuring how life events affect happiness. *International Journal of Epidemiology*, 31(6), pp.1139-1144.

Clark, A., Frijters, P. and Shields, M., 2008. Relative Income, Happiness, and Utility: An Explanation for the Easterlin Paradox and Other Puzzles. *Journal of Economic Literature*, 46(1), pp.95-144.

Clarke, K., 2007. A Simple Distribution-Free Test for Nonnested Model Selection. *Political Analysis*, 15(3), pp.347-363.

Czado, C., 2020. *Analyzing Dependent Data with Vine Copulas: A Practical Guide With R*. Springer.

Davillas, A. and Pudney, S., 2017. Concordance of health states in couples: Analysis of self-reported, nurse administered and blood-based biomarker data in the UK Understanding Society panel. *Journal of Health Economics*, 56, pp.87-102.

Davillas, A. and Pudney, S., 2020. Using biomarkers to predict healthcare costs: Evidence from a UK household panel. *Journal of Health Economics*, 73, p.102356.

Decancq, K., 2013. Copula-based measurement of dependence between dimensions of well-being. *Oxford Economic Papers*, 66(3), pp.681-701.

Denuit, M. and Lambert, P., 2005. Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis*, 93(1), pp.40-57.

Di Tella, R., Haisken-De New, J. and MacCulloch, R., 2010. Happiness adaptation to income and to status in an individual panel. *Journal of Economic Behavior & Organization*, 76(3), pp.834-852.

Di Tella, R., MacCulloch, R. and Oswald, A., 2003. The Macroeconomics of Happiness. *Review of Economics and Statistics*, 85(4), pp.809-827.

Dißmann, J., Brechmann, E., Czado, C. and Kurowicka, D., 2013. Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis*, 59, pp.52-69.

Ekman, P., Davidson, R. and Friesen, W., 1990. The Duchenne smile: Emotional expression and brain physiology: II. *Journal of Personality and Social Psychology*, 58(2), pp.342-353.

Ferrer-i-Carbonell, A. and Frijters, P., 2004. How Important is Methodology for the Estimates of the Determinants of Happiness?. *The Economic Journal*, 114(497), pp.641-659.

Ferrer-i-Carbonell, A. and Van Praag, B., 2008. [online] Available at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.319.7854&rep=rep1&type=pdf.

Ferrer-i-Carbonell, A., 2005. Income and well-being: an empirical analysis of the comparison income effect. *Journal of Public Economics*, 89(5-6), pp.997-1019.

Ferrer-i-Carbonell, A., 2013. Happiness economics. *SERIEs*, 4(1), pp.35-60.

Frijters, P. and Beatton, T., 2012. The mystery of the U-shaped relationship between happiness and age. *Journal of Economic Behavior & Organization*, 82(2-3), pp.525-542.

Gabriel, S., Mattey, J. and Wascher, W., 2003. Compensating differentials and evolution in the quality-of-life among U.S. states. *Regional Science and Urban Economics*, 33(5), pp.619-649.

Genest, C. and Favre, A., 2007. Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask. *Journal of Hydrologic Engineering*, 12(4), pp.347-368.

Gerdtham, U. and Johannesson, M., 2001. The relationship between happiness, health, and socio-economic factors: results based on Swedish microdata. *The Journal of Socio-Economics*, 30(6), pp.553-557.

Glenn, N., 2009. Is the apparent U-shape of well-being over the life course a result of inappropriate use of control variables? A commentary on Blanchflower and Oswald (66: 8, 2008, 1733–1749). *Social Science & Medicine*, 69(4), pp.481-485.

Goldberg, L., 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6), pp.1216-1229.

Goracci, A., Fagiolini, A., Martinucci, M., Calossi, S., Rossi, S., Santomauro, T., Mazzi, A., Penza, F., Fossi, A., Bargagli, E., Pieroni, M., Rottoli, P. and Castrogiovanni, P., 2008. Quality of life, anxiety and depression in Sarcoidosis. *General Hospital Psychiatry*, 30(5), pp.441-445.

Goyal, S., 2007. *Connections : An Introduction to the Economics of Networks*. Princeton: Princeton University Press.

Hamer, M. and Chida, Y., 2011. Life satisfaction and inflammatory biomarkers: The 2008 Scottish Health Survey1. *Japanese Psychological Research*, 53(2), pp.133-139.

Hernández-Alava, M. and Pudney, S., 2016. Bicop: A Command for Fitting Bivariate Ordinal Regressions with Residual Dependence Characterized by a Copula Function and Normal Mixture Marginals. *The Stata Journal: Promoting communications on statistics and Stata*, 16(1), pp.159-184.

Hofert, M., Kojadinovic, I., Machler, M. and Yan, J., 2018. *Elements of Copula Modeling with R*. Springer.

Hutson, A., Wilding, G., Mashtare, T. and Vexler, A., 2015. Measures of biomarker dependence using a copula-based multivariate epsilon–skew–normal family of distributions. *Journal of Applied Statistics*, 42(12), pp.2734-2753.

Joe, H., 1996. Families of m-Variate Distributions with Given Margins and m(m-1)/2 Bivariate Dependence Parameters. *Distributions with Fixed Marginals and Related Topics*, 28, pp.120-141.

Kim, J., Ju, H. and Jung, Y., 2020. Copula Approach for Developing a Biomarker Panel for Prediction of Dengue Hemorrhagic Fever. *Annals of Data Science*, 7(4), pp.697-712.

Kurowicka, D. and Cooke, R., 2006. *Uncertainty analysis with high dimensional dependence modelling*. Chichester: John Wiley & Sons.

Manfredini, R., Caraccioli, S., Salmi, R., Boeri, B., Tomelli, A. and Galrani, M., 2000. The Association of Low Serum Cholesterol with Depression and Suicidal Behaviours: New Hypotheses for the Missing Link. *Journal of International Medical Research*, 28(6), pp.247-257.

McCrae, R. and John, O., 1992. An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, 60(2), pp.175-215.

McFall, S., Petersen, J., Kaminska, O. and Lynn, P., 2014. *Understanding Society: The UK Household Longitudinal Study: Waves 2 and 3 Nurse Health Assessment, 2010 – 2012: Guide to Nurse Health Assessment*. Colchester: University of Essex.

Mejdoub, H. and Ben Arab, M., 2018. Impact of dependence modeling of non-life insurance risks on capital requirement: D-Vine Copula approach. *Research in International Business and Finance*, 45, pp.208-218.

Morales-Nápoles, O., 2011. Counting vines. In: D. Kurowicka and H. Joe, *Dependence modeling: vine copula handbook*. Singapore: World Scientific Publishing Co.

Oecd-ilibrary.org. 2011. *Divided We Stand*. [online] Available at: https://www.oecd-ilibrary.org/social-issues-migration-health/the-causes-of-growing-inequalities-in-oecd-countries_9789264119536-en.

Oswald, A. and Powdthavee, N., 2008. Does happiness adapt? A longitudinal study of disability with implications for economists and judges. *Journal of Public Economics*, 92(5), pp.1061-1077.

Oswald, A. and Wu, S., 2010. Objective Confirmation of Subjective Measures of Human Well-Being: Evidence from the U.S.A. *Science*, 327(5965), pp.576-579.

Pearson, T., Mensah, G., Alexander, R., Anderson, J., Cannon, R., Criqui, M., Fadl, Y., Fortmann, S., Hong, Y., Myers, G., Rifai, N., Smith, S., Taubert, K., Tracy, R. and Vinicor, F., 2003. Markers of Inflammation and Cardiovascular Disease. *Circulation*, 107(3), pp.499-511.

Pfaff, T., 2013. Income Comparisons, Income Adaptation, and Life Satisfaction: How Robust are Estimates from Survey Data?. *SSRN Electronic Journal*.

Poole, L., Hackett, R., Panagi, L. and Steptoe, A., 2019. Subjective wellbeing as a determinant of glycated haemoglobin in older adults: longitudinal findings from the English Longitudinal Study of Ageing. *Psychological Medicine*, 50(11), pp.1820-1828.

Prinsloo, S., Wei, Q., Scott, S., Tannir, N., Jonasch, E., Pisters, L. and Cohen, L., 2015. Psychological states, serum markers and survival: associations and predictors of survival in patients with renal cell carcinoma. *Journal of Behavioral Medicine*, 38(1), pp.48-56.

Quinn, C., 2007a. The health-economic applications of copulas: methods in applied econometric research. *HEDG Working Paper 07/22*.

Quinn, C., 2007b. Using copulas to measure association between ordinal measures of health and income. *HEDG Working Paper 07/24*.

Radler, B., Rigotti, A. and Ryff, C., 2018. Persistently high psychological well-being predicts better HDL cholesterol and triglyceride levels: findings from the midlife in the U.S. (MIDUS) longitudinal study. *Lipids in Health and Disease*, 17(1).

Rakonczai, P., Rojkovich, B. and Gáti, T., 2015. Conditional Copula Models With Applications to Biomarkers In Rheumatoid Arthritis. *Value in Health*, 18(7), pp.A705-A706.

Romero Martinez, V., Silva, E. and Villasmil, J., 2010. RELATIONSHIP BETWEEN LIFE SATISFACTION LEVELS AND HIGH BLOOD PRESSURE IN ADOLESCENTS: PP.14.43. *Journal of Hypertension*, 28, p.e260.

Sandvik, E., Diener, E. and Seidlitz, L., 1993. Subjective Well-Being: The Convergence and Stability of Self-Report and Non-Self-Report Measures. *Journal of Personality*, 61(3), pp.317-342.

Schenk, H., Jeronimus, B., van der Krieke, L., Bos, E., de Jonge, P. and Rosmalen, J., 2018. Associations of Positive Affect and Negative Affect With Allostatic Load: A Lifelines Cohort Study. *Psychosomatic Medicine*, 80(2), pp.160-166.

Senik, C., 2004. When information dominates comparison: Learning from Russian subjective panel data. *Journal of Public Economics*, 88(9), pp.2099-2123.

Sklar, A., 1959. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris*, 8, pp.229-231.

Sutton, S. and Davidson, R., 1997. Prefrontal Brain Asymmetry: A Biological Substrate of the Behavioral Approach and Inhibition Systems. *Psychological Science*, 8(3), pp.204-210.

Szabo, A., Böhm, T. and Köteles, F., 2020. Relationship between aerobic fitness, blood pressure and life satisfaction. *Baltic Journal of Health and Physical Activity*, 12(2), pp.1-11.

Tsenkova, V., Dienberg Love, G., Singer, B. and Ryff, C., 2008. Coping and positive affect predict longitudinal change in glycosylated haemoglobin. *Health Psychology*, 27(2), pp.S163-S171.

Valtysdottir, S., Wide, L. and Hallgren, R., 2003. Mental wellbeing and quality of sexual life in women with primary Sjögren's syndrome are related to circulating dehydroepiandrosterone sulphate. *Annals of the Rheumatic Diseases*, 62(9), pp.875-879.

Vuong, Q., 1989. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, 57(2), pp.307-333.

WHO, 2011. [online] Ncbi.nlm.nih.gov. Available at: https://www.ncbi.nlm.nih.gov/books/NBK304267/pdf/Bookshelf_NBK304267.pdf.

Wong, S., Leung, J., Kwok, T., Ohlsson, C., Vandenput, L., Leung, P. and Woo, J., 2011. Low DHEAS levels are associated with depressive symptoms in elderly Chinese men: results from a large study. *Asian Journal of Andrology*, 13(6), pp.898-902.

Wood, A., Boyce, C., Moore, S. and Brown, G., 2012. An evolutionary based social rank explanation of why low income predicts mental distress: A 17 year cohort study of 30,000 people. *Journal of Affective Disorders*, 136(3), pp.882-888.

Yoo, J., Miyamoto, Y. and Ryff, C., 2016. Positive affect, social connectedness, and healthy biomarkers in Japan and the U.S. *Emotion*, 16(8), pp.1137-1146.

Zhang, D., Yan, M. and Tsopanakis, A., 2018. Financial stress relationships among Euro area countries: an R-vine copula approach. *The European Journal of Finance*, 24(17), pp.1587-1608.

APPENDIX A

A.1 Fundamentals

Loosely speaking, copula theory facilitates the examination of the dependence between the elements of a set of random variables in a manner that permits studying the dependence structure as a separate entity from each of the univariate marginal distributions of the variables in the set. The main takeaway from the current subsection is the fact that this dependence structure can be represented by a function classified as a copula, which under certain conditions is unique for a set of variables. The presentation in this subsection and subsequent ones is based on the expositions by Hofert *et al.* (2018), and Czado (2020).

A d-dimensional copula $C(u_1, \dots, u_d)$ where $(u_1, \dots, u_d) \in [0,1]^d$ is a multivariate distribution function defined on the unit hypercube. The univariate margins of the copula are uniformly distributed on $[0,1]$. For an absolutely continuous copula[56], the copula density $c(u_1, \dots, u_d)$ can be obtained by:

$c(u_1, \dots, u_d) = \frac{\partial^d}{\partial u_1 \dots \partial u_d} C(u_1, \dots, u_d)$ for all $(u_1, \dots, u_d) \in [0,1]^d$.

The usefulness of copulas is apparent through Sklar's theorem. According to Sklar (1959), for a d-dimensional vector of random variables $\boldsymbol{X} = (X_1, \dots, X_d)^T$ with joint distribution function $F_{\boldsymbol{X}}(x_1, \dots, x_d)$ where $(x_1, \dots, x_d) \in \mathbb{R}^d$, and marginal distribution functions $F_i(x_i)$ where $x_i \in \mathbb{R}$, and $i \in \{1, \dots, d\}$, the joint distribution function can be expressed as:

$F_{\boldsymbol{X}}(x_1, \dots, x_d) = C_{\boldsymbol{X}}\big(F_1(x_1), \dots, F_d(x_d)\big)$ where $(x_1, \dots, x_d) \in \mathbb{R}^d$.

$C_{\boldsymbol{X}}(u_1, \dots, u_d)$ where $(u_1, \dots, u_d) \in [0,1]^d$ is a copula.

The relevant probability density function[57] $f_{\boldsymbol{X}}(x_1, \dots, x_d)$ can be expressed as:

$f_{\boldsymbol{X}}(x_1, \dots, x_d) = c_{\boldsymbol{X}}\big(F_1(x_1), \dots, F_d(x_d)\big) f_1(x_1) \dots f_d(x_d)$ where $(x_1, \dots, x_d) \in \prod_{i=1}^{d} \mathrm{ran}X_i$,

where $f_i(x_i)$ for $x_i \in \mathbb{R}$, and $i \in \{1, \dots, d\}$ represents the probability density function for variable X_i. For absolutely continuous marginal distribution functions $\mathrm{ran}X_i = \{x \in \mathbb{R}: f_i(x) > 0\}$ for $i \in \{1, \dots, d\}$. $c_{\boldsymbol{X}}(u_1, \dots, u_d)$ where $(u_1, \dots, u_d) \in [0,1]^d$ is the density of

[56] An absolutely continuous copula admits a density (Hofert *et al.*, 2018).

[57] From Sklar's theorem, a joint distribution function is absolutely continuous if and only if both the copula and the marginal distribution functions are absolutely continuous (Hofert *et al.*, 2018). Note that in practice most of the continuous distribution functions used in applied statistics are also absolutely continuous. Therefore, the distinction between continuity and absolute continuity is more significant theoretically rather than practically for this paper.

$C_{\boldsymbol{X}}(u_1, \dots, u_d)$. For absolutely continuous joint distribution functions, the copula is unique (Czado, 2020)[58].

Loosely speaking, Sklar's theorem implies that the study of the joint distribution function for a set of random variables can be decomposed into the study of the dependence structure as captured by the copula function, and the separate study of the marginal distribution functions.

A.2 Invariance of copulas

For a vector of random variables $\boldsymbol{X} = (X_1, \dots, X_d)^T$ with absolutely continuous marginal distribution functions, the unique copula as implied by Sklar's theorem is invariant to strictly increasing transformations of the individual variables. More formally, for $Y_i = T_i(X_i)$ for $i \in \{1, \dots, d\}$, and T_i representing a strictly increasing function, the unique copula of $\boldsymbol{Y} = (Y_1, \dots, Y_d)^T$ denoted by $C_{\boldsymbol{Y}}$, as implied by Sklar's theorem, is identical to the unique copula $C_{\boldsymbol{X}}$ of the underlying vector of random variables $\boldsymbol{X}$. Proof for the invariance of copulas is provided in subsection 1.7 of Czado (2020).

Loosely speaking, any strict monotonic transformation of the elements in the vector of random variables considered will not alter the dependence structure as represented by the copula. In this sense, copulas allow modelling the association between variables of interest in a 'margin-free' way.

A.3 Rotated copulas

Certain parametric copula families, such as the Gumbel and Joe copula families, can only be used to capture the dependence structure between variables which have a positive association. Optimally, it should be the case that negative association can be captured as well by any of the parametric copula families used. To overcome this issue for the bivariate case, counterclockwise rotations of the copulas for parametric families which present this restriction can be used. For a copula $C(u_1, u_2)$ where $(u_1, u_2) \in [0,1]^2$ with density $c(u_1, u_2)$ the rotated versions are such that $C_{90}(u_1, u_2) = u_2 - C(1 - u_1, u_2)$ for a 90 degrees rotation, $C_{180}(u_1, u_2) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2)$ for a 180 degrees rotation, and $C_{270}(u_1, u_2) = u_1 - C(u_1, 1 - u_2)$ for a 270 degrees rotation. The corresponding copula densities are such that $c_{90}(u_1, u_2) = c(1 - u_1, u_2)$ for a 90 degrees rotation, $c_{180}(u_1, u_2) = c(1 - u_1, 1 - u_2)$ for a 180 degrees rotation, and $c_{270}(u_1, u_2) = c(u_1, 1 - u_2)$ for a 270

[58] Sklar's theorem is based on the notion of the probability integral transform. For a random variable with an absolutely continuous distribution function $X \sim F$, the transformation $U = F(X)$, known as the probability integral transform, is uniformly distributed on [0,1]. A proof is provided in subsection B.4.

degrees rotation. In addition to overcoming the aforementioned issue, using the rotated versions of the set of parametric copula families considered enhances the flexibility in modelling bivariate dependence.

A.4 Bivariate conditional distributions and *h-functions*

For a 2-dimensional vector of random variables $\boldsymbol{X} = (X_1, X_2)^T$ with absolutely continuous joint distribution function F such that $F(x_1, x_2) = C_{12}\big(F_1(x_1), F_2(x_2)\big)$ for a unique copula C_{12} where $(x_1, x_2) \in \mathbb{R}^2$, the conditional distribution function $F_{1|2}(x_1|x_2)$ and the conditional density function $f_{1|2}(x_1|x_2)$ can be given as

$F_{1|2}(x_1|x_2) = \frac{\partial}{\partial F_2(x_2)} C_{12}\big(F_1(x_1), F_2(x_2)\big)$,

and $f_{1|2}(x_1|x_2) = c_{12}\big(F_1(x_1), F_2(x_2)\big) f_1(x_1)$.

Proofs are provided in subsection 1.9 of Czado (2020).

Given that the 2-dimensional (absolutely continuous) copula $C_{12}(u_1, u_2)$ where $(u_1, u_2) \in [0,1]^2$ is a bivariate distribution function, this implies that

$C_{1|2}(u_1|u_2) = \frac{\partial}{\partial u_2} C_{12}(u_1, u_2)$ for all $(u_1, u_2) \in [0,1]^2$.

This conditional distribution function is denoted by Aas *et al.* (2009) as an *h-function* such that

$h_{1|2}(u_1|u_2) = \frac{\partial}{\partial u_2} C_{12}(u_1, u_2)$,

and $h_{2|1}(u_2|u_1) = \frac{\partial}{\partial u_1} C_{12}(u_1, u_2)$.

Furthermore, based on the aforementioned expressions, it is implied that

$F_{1|2}(x_1|x_2) = C_{1|2}\big(F_1(x_1)|F_2(x_2)\big) = h_{1|2}\big(F_1(x_1)|F_2(x_2)\big)$.

The usefulness of *h-functions* will become apparent further on in the specification and estimation steps of the vine copula model as it is used in subsection B.3. The notion of the *h-function* is used in a recursive manner to generate conditional distributions where the conditioning set of variables is not necessarily a singleton as in the example mentioned in this subsection. In the aforementioned example the conditioning set is given by $\{X_1\}$ or $\{X_2\}$ depending on the specific *h-function* derived.

A.5 Dimensionality and vine copulas

In the bivariate case, copulas offer great flexibility in modelling joint distribution functions. Several parametric copula classes exist, such as the Elliptical and Archimedean classes, which nest multiple families of copulas themselves (e.g. the Gumbel and Joe copula families for the class of Archimedean copulas). Different copula families can be used to capture features such as asymmetry and (asymmetric) tail dependence in modelling a bivariate distribution function, as opposed to the case of using e.g. the normal distribution.

However, the same level of flexibility is not necessarily true for a higher number of dimensions since parametric copula families are not as well-investigated as they are for the bivariate case (Brechmann and Schepsmeier, 2013). Vine copulas or pair copula constructions can be used to circumvent the issue of flexibility in a higher number of dimensions (Joe, 1996). A vine copula represents a specific decomposition of the multivariate probability density function into bivariate copula densities through repeated conditioning. Each bivariate copula can be chosen independently offering high flexibility in modelling. A more elaborate presentation of vine copula modelling is offered in the next section.

APPENDIX B

B.1 Pair copula decompositions

A pair copula decomposition refers to the representation of a multivariate probability density function in terms of the corresponding univariate marginal density functions, and (conditional) bivariate copula densities. This decomposition is achieved through repeated conditioning, and thus it is not unique for a particular joint density function.

For illustrative purposes, a pair copula decomposition of the 3-dimensional probability density function $f_{123}(x_1, x_2, x_3)$ where $(x_1, x_2, x_3) \in \mathbb{R}^3$ is considered. $F_i(x_i)$ and $f_i(x_i)$ represent the distribution and density functions respectively for variable $X_i, i \in \{1,2,3\}$[59]. Based on repeated conditioning,

$$f_{123}(x_1, x_2, x_3) = f_{3|12}(x_3|x_1, x_2) f_{2|1}(x_2|x_1) f_1(x_1).$$

By using the equalities presented in subsection A.4,

$$f_{2|1}(x_2|x_1) = c_{12}\big(F_1(x_1), F_2(x_2)\big) f_2(x_2).$$

Furthermore, it is also implied that,

$$f_{3|12}(x_3|x_1, x_2) = c_{13;2}\big(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2\big) f_{3|2}(x_3|x_2)$$[60],

where $f_{3|2}(x_3|x_2) = c_{23}\big(F_2(x_2), F_3(x_3)\big) f_3(x_3)$.

Therefore, putting it all together,

$$f_{123}(x_1, x_2, x_3) = c_{13;2}\big(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2\big) \dots$$

$$\times c_{12}\big(F_1(x_1), F_2(x_2)\big) c_{23}\big(F_2(x_2), F_3(x_3)\big) \dots$$

$$\times f_1(x_1) f_2(x_2) f_3(x_3).$$

The 3-dimensional probability density function is therefore decomposed into a conditional bivariate copula density, two unconditional bivariate copula densities, and the univariate marginal density functions.

[59] Again, using mainly the exposition by Czado (2020).

[60] Note that, in general, $C_{13;2}$ is not equal to $C_{13|2}$. $C_{13;2}$ is used to denote a copula with standard uniform margins, whereas $C_{13|2}$ is used to denote a bivariate conditional distribution function which is not necessarily a copula. This is true for any pair of conditioned variables and any set of conditioning variables. $C_{13;2}$ is referred to as a conditional bivariate copula.

Note that the conditional bivariate copula density in the 3-dimensional example given depends on the value of the conditioning variable, x_2 in this case. To be able to use the notion of the pair copula decomposition in a constructive manner, this dependence on the conditioning variable(s) is usually ignored. This is known as the simplifying assumption. More formally, for the example considered above,

$$c_{13;2}\big(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2); x_2\big) = c_{13;2}\left(F_{1|2}(x_1|x_2), F_{3|2}(x_3|x_2)\right) \text{ for all } x_2 \in \mathbb{R}.$$

Given that the simplifying assumption holds, arbitrary choices of pair copulas $c_{13;2}$, c_{12}, and c_{23} in the example above can be made independently from the set of parametric copulas, such that, along with the marginal density functions $f_1(x_1)$, $f_2(x_2)$, and $f_3(x_3)$, they constitute the building blocks for the construction of a parametric trivariate density function. This is known as a pair copula construction. The logic of the simplifying assumption and the pair copula construction extends to any multivariate joint density function of an arbitrary number of dimensions d.

B.2 Regular vines

As seen in the previous subsection, a multivariate density function can be decomposed into (conditional) bivariate copula densities. Given the simplifying assumption, an arbitrary parametric copula family can be assigned to each of these bivariate copulas such that they act as building blocks for the construction of a parametric multivariate density function[61]. As noted, any decomposition is not unique. Depending on the approach of iterative conditioning followed many different pair copula constructions can arise for a particular set of variables.

The set of pair copula constructions considered in this paper belongs to a subclass of vine copulas known as regular vine copulas. Such pair copula constructions are coupled with a graphical structure, known as a regular vine, which captures the particular decomposition considered (Bedford and Cooke, 2001). In order to use the graphical representation, some fundamental concepts of graph theory are introduced. An extensive review of graph theory is provided in Goyal (2007).

[61] Regarding the univariate marginal density functions which are part of the pair copula decomposition, a non-parametric estimation procedure is followed to transform the data into what is known as pseudo-copula data. Individual variables from the pseudo-copula data are assumed to follow a standard uniform distribution. For a variable $X \sim U[0,1]$, the probability density function $f(x) = 1$ for all $x \in [0,1]$. Therefore, assigning a parametric copula family to each of the (conditional) bivariate copula densities suffices for a parametric pair copula construction. A more detailed description will be provided further on.

A graph $G = (N, E)$ consists of a finite set of n nodes, $N = \{1, \dots, n\}$, and the set of edges or links between them E. The existence of an edge is depicted by the binary variable $g_{ij} \in \{0,1\}$ which takes the value of 1 if a direct relationship exists between nodes i and j, and 0 otherwise. For the purposes of this paper only undirected links are considered, i.e. $g_{ij} = g_{ji}$ for all $(i, j) \in \{1, \dots, n\}^2$. The set of edges E consists of all the pairs of nodes with a direct relationship between them, such that $E = \{(i, j) \in \{1, \dots, n\}^2 : g_{ij} = 1\}$[62].

Loosely speaking, the set of nodes for a graph which is classified as a regular vine consists of all the arguments of the (conditional) bivariate copula densities which make up the chosen pair copula construction. If a link is present between two nodes then the direct relationship represented is in the form of the fact that the two arguments represented by the linked nodes are both arguments of a particular (conditional) bivariate copula density. Before proceeding to a more detailed definition of the regular vine, some additional concepts linked to graphs are presented which are used at various points along the path to model estimation.

Given any graph $G = (N, E)$, there exists a path between two distinct nodes i and j either if $g_{ij} = 1$, or if there is a set of distinct intermediate nodes $\{j_1, j_2, \dots, j_n\} \subseteq N$ such that $g_{ij_1} = g_{j_1 j_2} = \cdots = g_{j_n j} = 1$. Therefore, if a path exists between two nodes then this implies that they are either directly or indirectly connected. A connected graph is such that there is a path between any pair of nodes $(i, j) \in \{1, \dots, n\}^2$. A graph $G = (N, E)$ contains a cycle if there exists a set of distinct nodes $\{j_1, j_2, \dots, j_n\} \subseteq N$ such that $g_{j_1 j_2} = \cdots = g_{j_n j_1} = 1$, i.e. there exists a path which connects a node in the set $\{j_1, j_2, \dots, j_n\} \subseteq N$ to itself. A graph which does not contain a cycle is known as acyclic. Lastly, a graph is complete if $g_{ij} = 1$ for all $(i, j) \in \{1, \dots, n\}^2$, i.e. there exists a direct relationship between any pair of nodes of the graph.

One type of graph is of particular interest when considering regular vines. This type of graph is called a tree. A tree is defined as an undirected, connected, acyclic graph or, equivalently, an undirected graph in which any pair of nodes is connected by a unique path. The property of being undirected for the graph refers to the existence of only undirected links. Furthermore, for a graph $G = (N, E)$, a subgraph $\bar{G} = (\bar{N}, \bar{E})$ is such that $\bar{N} \subseteq N$ and $\bar{E} \subseteq E$, i.e. a subgraph of a particular graph is a graph with only a subset of the nodes and edges of the original graph.

[62] Note that $g_{ii} = 0$ for all $i \in \{1, \dots, n\}$.

Lastly, for a graph $G = (N, E)$, a spanning tree $\bar{G} = (\bar{N}, \bar{E})$ is a subgraph of G such that $\bar{G}$ satisfies the definition of a tree with $\bar{N} = N$.

The graphical structure which corresponds to the representation of a regular vine copula is a sequence of trees which satisfies a set of assumptions, and where each distinct tree in the sequence satisfies the aforementioned graph theoretical definition for a tree. Czado (2020) presents the set of assumptions which suffice for a sequence of trees $\{T_1, \dots, T_{d-1}\}$ to constitute a regular vine tree sequence that represents a pair copula construction for a d-dimensional density function. In particular, T_1 is defined by a set of nodes $N_1 = \{1, \dots, d\}$ and a set of edges E_1. Furthermore, T_k for $k \in \{2, \dots, d-1\}$ is defined by a set of nodes $N_k = E_{k-1}$ and a set of edges E_k. Loosely speaking, this suggests that the edges in any tree become the nodes of the subsequent tree. Lastly, if there is an edge connecting two nodes in tree T_k for $k \in \{2, \dots, d-1\}$, then these nodes which represent edges in tree T_{k-1} must share a common node in tree T_{k-1}, i.e. there is a node in tree T_{k-1} on which both edges are attached[63;64].

For the first tree T_1 of a regular vine tree sequence in d dimensions the set of nodes can be given by $N_1 = \{1, \dots, d\}$. The set of edges E_1 is such that $E_1 = \{(i,j) \in \{1, \dots, n\}^2 : g_{ij} = 1\}$. Therefore, based on the aforementioned assumption for a regular vine, the set of nodes for T_2 is such that $N_2 = E_1$. For tree T_k where $k \in \{2, \dots, d-1\}$ the set of edges E_k is denoted by $E_k = \left\{\left(i, j, D_i, D_j\right) : g_{ij} = 1, (i,j) \in N_k^2 = E_{k-1}^2, D_i \subseteq \{1, \dots, d\}, D_j \subseteq \{1, \dots, d\}\right\}$. This suggests that for trees in a regular vine, after the first one, each existing edge is defined not only by the pair of nodes which are linked in the tree, but also by a pair of sets D_i and D_j which represent subsets of the set of nodes defining the first tree. As shown in subsection B.3, the set of nodes for the first tree basically represents the original vector of random variables under examination. As such, each of the sets D_i and D_j essentially represents a subset of a vector of random variables. More details are provided in the next subsection.

For node $i \in N_k = E_{k-1}$ the set $D_i = \{j \in N_1 : \exists i_1 \in E_1, \dots, i_{k-2} \in E_{k-2} \; such \; that \; j \in i_1 \in \cdots \in i_{k-2} \in i\}$. Loosely speaking, this suggests that the set D_i for node i consists of all those elements in the set of nodes which define the first tree of the sequence that are used in a sequential composition of edges leading up to i. Given that $g_{ij} = 1$, the sets A_{ij} and B_{ij} for

[63] This is known as the proximity condition.

[64] It is worth noting that there are $\left(\frac{d!}{2}\right) \times 2^{\binom{d-2}{2}}$ possible regular vine tree sequences which satisfy the assumptions for d dimensions (Morales-Nápoles, 2011).

$(i,j) \in N_k^2 = E_{k-1}^2$ are such that $A_{ij} = D_i \cap D_j$ and $B_{ij} = H_i \cup H_j$ where $H_i = \{D_i \backslash A_{ij}\}$. It can be shown that the set B_{ij} contains two distinct elements for all $(i,j) \in N_k^2 = E_{k-1}^2$ and $k \in \{2, \dots, d-1\}$ (Kurowicka and Cooke, 2006). For $(i,j) \in N_1^2$ such that $g_{ij} = 1$, $A_{ij} = \{\emptyset\}$ and $B_{ij} = \{i,j\}$.

The necessity of this specification will become apparent in subsection B.3 where the regular vine tree sequence is coupled with the regular vine copula, such that the set of nodes of the first tree N_1 represents the d-dimensional vector of the variables of interest. The existence of an edge depicts the existence of a bivariate copula density such that the linked nodes determine the arguments of the density. For each existing edge between nodes i and j the set B_{ij} represents the arguments of the density associated to the edge, and the set A_{ij} represents the set of conditioning variables for a conditional bivariate copula density. A detailed description follows.

B.3 Regular vine copulas

Based on the concepts introduced in subsections B.1 and B.2, the notion of a regular vine copula is presented in this subsection. Extensive reviews are provided by Kurowicka and Cooke (2006), and Czado (2020). Loosely speaking, this notion refers to the coupling of a particular pair copula construction for a multivariate density function with a regular vine tree sequence which satisfies the assumptions presented in the previous subsection.

For a d-dimensional density function any pair copula decomposition, and thus any pair copula construction, is not unique. However, the end result of any pair copula decomposition consists of the product of the d corresponding univariate marginal density functions and a set of (conditional) bivariate copula densities as it can be seen in subsection B.1 for the 3-dimensional case.

In order to present the notion of a regular vine copula, the more general notion of a regular vine distribution is introduced first. A regular vine distribution is characterised by three components, a set of univariate distribution functions $\mathcal{F}$, a regular vine tree sequence, and a set of bivariate copulas $\mathcal{B}$. The d-dimensional vector of random variables $\boldsymbol{X} = (X_1, \dots, X_d)^T$ with absolutely continuous joint distribution function $F_{\boldsymbol{X}}(x_1, \dots, x_d)$ where $(x_1, \dots, x_d) \in \mathbb{R}^d$, and marginal distribution functions $F_i(x_i)$ where $x_i \in \mathbb{R}$ and $i \in \{1, \dots, d\}$ is said to have a regular vine distribution if the distribution of $\boldsymbol{X}$ can be characterised by the three aforementioned components as follows. The set $\mathcal{F}$ is such that $\mathcal{F} = \{F_1(x_1), \dots, F_d(x_d)\}$, i.e. the set $\mathcal{F}$ is

composed of the univariate marginal distribution functions of vector $\boldsymbol{X}$. The regular vine tree sequence consists of a sequence of trees $\{T_1, \dots, T_{d-1}\}$, where $T_k = (N_k, E_k)$ for $k \in \{1, \dots, d-1\}$, which satisfies the assumptions presented by Czado (2020), and described in subsection B.2. The set of nodes N_1 for the first tree of the sequence T_1 is assumed to represent the vector of random variables $\boldsymbol{X}$. Lastly, the set of bivariate copulas $\mathcal{B}$ consists of symmetric bivariate copulas which admit a density function[65]. What completes the characterisation of the distribution of $\boldsymbol{X}$ as a regular vine distribution is the specification of the relationship between the set $\mathcal{B}$ and the regular vine tree sequence.

There is a link between the structure of the regular vine tree sequence and the set of bivariate copulas. In particular, for each $(i,j) \in N_k^2$ where $k \in \{1, \dots, d-1\}$ such that $g_{ij} = 1$, i.e. for all existing edges in the sequence, there exists a copula $C_{ij} \in \mathcal{B}$ which is associated (according to Sklar's theorem) to the joint conditional distribution of the variables corresponding to the elements in the set B_{ij} given the variables corresponding to the elements of the set A_{ij}. c_{ij} is used to denote the copula density of C_{ij}. The sets A_{ij}, B_{ij}, and H_i are as defined in subsection B.2. Therefore, given that N_1 is assumed to represent the vector of random variables $\boldsymbol{X}$, the elements of sets A_{ij}, B_{ij}, and H_i also represent individual variables from the vector $\boldsymbol{X}$[66]. Therefore, $\boldsymbol{X}_{A_{ij}}$, $\boldsymbol{X}_{B_{ij}}$, and $\boldsymbol{X}_{H_i}$ are used to denote the variables which correspond to the elements of sets A_{ij}, B_{ij}, and H_i[67].

Given a set $\mathcal{F}$ with size d, a regular vine tree sequence $\{T_1, \dots, T_{d-1}\}$, and a set of bivariate copulas $\mathcal{B}$ where all three components satisfy the aforementioned description, a d-dimensional density function $f_{\boldsymbol{X}}(x_1, \dots, x_d)$ where $(x_1, \dots, x_d) \in \mathbb{R}^d$ can be identified such that[68]

$$f_{\boldsymbol{X}}(x_1, \dots, x_d) = \prod_{k=1}^{d-1} \prod_{\{(i,j) \in N_k^2 : g_{ij}=1\}} c_{ij}\left(F_{\boldsymbol{X}_{H_i}|\boldsymbol{X}_{A_{ij}}}\left(\boldsymbol{x}_{H_i}|\boldsymbol{x}_{A_{ij}}\right), F_{\boldsymbol{X}_{H_j}|\boldsymbol{X}_{A_{ij}}}\left(\boldsymbol{x}_{H_j}|\boldsymbol{x}_{A_{ij}}\right)\right)\dots$$

$$\times \prod_{i=1}^{d} f_i(x_i).$$

[65] A bivariate copula C with density c is called symmetric or exchangeable in this case if $c(u_1, u_2) = c(u_2, u_1)$ for all $(u_1, u_2) \in [0,1]^2$. The symmetry condition is not necessary. Non-symmetric bivariate copulas can be used, accompanied by the adjustment that the regular vine tree sequence also includes directed links, i.e. $g_{ij} \neq g_{ji}$ for some $(i,j) \in N_k^2$ where $k \in \{1, \dots, d-1\}$.

[66] Recall that $|B_{ij}| = 2$ for all $(i,j) \in N_k^2$ and $k \in \{1, \dots, d-1\}$ such that $g_{ij} = 1$ which implies that only bivariate copulas are needed in set $\mathcal{B}$.

[67] Note that for the edges of the first tree the corresponding copulas in the set $\mathcal{B}$ are associated to joint unconditional distributions given that the set A_{ij} is empty for each $(i,j) \in N_1^2$ such that $g_{ij} = 1$.

[68] Note that for each particular triplet of $\mathcal{F}$, $\{T_1, \dots, T_{d-1}\}$, and $\mathcal{B}$ a unique multivariate density function is identified (Bedford and Cooke, 2002). This is true even in the case of using non-symmetric bivariate copulas.

c_{ij} refers to a (conditional) bivariate copula density corresponding to the copula $C_{ij} \in \mathcal{B}$. C_{ij} is such that

$$F_{\boldsymbol{X}_{B_{ij}}|\boldsymbol{X}_{A_{ij}}}\left(\boldsymbol{x}_{B_{ij}}|\boldsymbol{x}_{A_{ij}}\right) = C_{ij}\left(F_{\boldsymbol{X}_{H_i}|\boldsymbol{X}_{A_{ij}}}\left(\boldsymbol{x}_{H_i}|\boldsymbol{x}_{A_{ij}}\right), F_{\boldsymbol{X}_{H_j}|\boldsymbol{X}_{A_{ij}}}\left(\boldsymbol{x}_{H_j}|\boldsymbol{x}_{A_{ij}}\right)\right).$$

Note that the simplifying assumption is implicitly used as C_{ij} does not depend on $\boldsymbol{x}_{A_{ij}}$ for all $(i,j) \in N_k^2$ where $k \in \{1,\dots,d-1\}$ such that $g_{ij} = 1$. $f_i(x_i)$ for $i \in \{1,\dots,d\}$ represent the density functions for the elements of the set $\mathcal{F}$.

In the aforementioned exposition of the d-dimensional density function $f_{\boldsymbol{X}}$, the specification of the conditional distributions which act as arguments for conditional bivariate copula densities is also required. This is where the notion of the *h-function* presented in subsection A.4 becomes useful. The results presented in subsection A.4 can be generalised to any number of dimensions greater than 2. As such, for random variables X and Y, and a vector of random variables $\boldsymbol{Z}$ which all together have an absolutely continuous joint distribution function it is the case that

$$F_{X|Y,\boldsymbol{Z}}(x|y,\boldsymbol{z}) = \frac{\partial}{\partial F_{Y|\boldsymbol{Z}}(y|\boldsymbol{z})} C_{X,Y;\boldsymbol{Z}}\left(F_{X|\boldsymbol{Z}}(x|\boldsymbol{z}), F_{Y|\boldsymbol{Z}}(y|\boldsymbol{z})\right).$$

A recursive application of this rule can be used to specify the arguments of the conditional bivariate copula densities.

A regular vine copula is simply defined as a regular vine distribution where all elements of the set $\mathcal{F}$ represent the standard uniform distribution. For a variable $X \sim U[0,1]$, the probability density function $f(x) = 1$ for all $x \in [0,1]$. As such, in the specification of the multivariate density function identified for the regular vine distribution, the component which accounts for the product of the univariate marginal density functions is equal to 1.

B.4 Marginal density function specification

Before considering the model selection and estimation steps, the non-parametric approach based on which the univariate marginal density functions in the specification of a regular vine distribution are dealt with is presented.

This is based on the notion of the probability integral transform. In particular, for a random variable X with an absolutely continuous distribution function F, the transformation $U = F(X)$,

known as the probability integral transform, is uniformly distributed on $[0,1]$. This is true since $P(U \leq u) = P(F(X) \leq u) = P\left(X \leq F^{-1}(u)\right) = F\left(F^{-1}(u)\right) = u$ is true for any $u \in [0,1]$[69].

As seen, for a vector of random variables $\boldsymbol{X} = (X_1, \dots, X_d)^T$ with absolutely continuous joint distribution function $F_{\boldsymbol{X}}$ and marginal distribution functions F_i for $i \in \{1, \dots, d\}$ there exists a unique copula which satisfies Sklar's theorem. An absolutely continuous joint distribution function implies absolutely continuous marginal distribution functions (and an absolutely continuous copula; Hofert *et al.*, 2018). As such, $\boldsymbol{Y} = \left(F_1(X_1), \dots, F_d(X_d)\right)$ basically represents a vector of strictly monotonic transformations of the individual variables in $\boldsymbol{X}$. Based on the concept of invariance of copulas presented in subsection A.2, the unique copula which satisfies Sklar's theorem for $\boldsymbol{Y}$ is equivalent to the one for $\boldsymbol{X}$. By the notion of the probability integral transform, $F_i(X_i)$ has a standard uniform distribution for each $i \in \{1, \dots, d\}$. As such, by considering $\boldsymbol{Y}$ instead of $\boldsymbol{X}$ a regular vine copula can be considered instead of a regular vine distribution while the characterisation of dependence between the variables as provided by the copula remains the same.

One possible option in applying the aforementioned transformation is to assume a parametric absolutely continuous marginal distribution function F_i for each variable X_i for $i \in \{1, \dots, d\}$. However, such an approach runs the risk of misspecification. Instead, a non-parametric method known as the empirical distribution function is used in this paper. In particular, given x_i for $i \in \{1, \dots, n\}$ independent and identically distributed observations of a random variable X with distribution function F, the empirical distribution function $\hat{F}$ is such that

$\hat{F}(x) = \frac{1}{n+1} \sum_{i=1}^{n} 1(x_i \leq x)$ for any $x \in \mathbb{R}$,

where $1(\cdot)$ is the indicator function.

The empirical distribution function is applied to each individual variable from the set of variables examined to transform the observed values for each variable to the so-called pseudo-copula data. In particular, for a d-dimensional vector of random variables $\boldsymbol{X}$ an $n \times d$ matrix of values $\boldsymbol{x}$ is observed where x_{ij} represents the observation from subject $i \in \{1, \dots, n\}$ for variable $j \in \{1, \dots, d\}$. The $n \times d$ matrix of pseudo-copula observations $\boldsymbol{u}$ is defined such that

[69] Note that F^{-1} in this proof represents a generalized inverse which takes the form of the standard inverse for strictly increasing distributions functions, or the form of the quantile function for increasing distribution functions.

$u_{ij} = \frac{1}{n+1}\sum_{k=1}^{n} 1\left(x_{kj} \leq x_{ij}\right)$ for all $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, d\}$[70].

Based on this approach, in combination with a parametric copula specification, the suitable estimation method is maximum pseudo-likelihood estimation (Hofert *et al.*, 2018). More on this will follow.

B.5 Model selection and estimation

As already seen, to fully specify a regular vine copula model three components are required, a set of univariate distribution functions, a set of bivariate copulas, and a regular vine tree sequence. Given a regular vine tree sequence, and parametric choices for the univariate distribution functions and the bivariate copulas, maximum likelihood estimation can be used to derive estimates for the parameters related to the assumed specification. However, for the purposes of this paper, the set of univariate distributions is estimated non-parametrically, as demonstrated in subsection B.4. A parametric specification for the set of bivariate copulas is implemented.

Despite the non-parametric estimation applied to the univariate distribution functions, the problem of model selection is still cumbersome. There is a vast amount of possible regular vine tree sequences and a large set of parametric copula families to choose from. The model selection strategy followed proceeds sequentially, tree by tree, where the structure for each tree is chosen starting from the first tree of the sequence, followed by the choice of the parametric bivariate copulas which correspond to the edges of the tree. The parameters associated to each tree are estimated before proceeding to structure selection for the subsequent tree of the regular vine tree sequence.

The first thing presented is the method based on which the set of bivariate copulas is selected. For a given regular vine tree sequence, a choice is made from the set of parametric copula families for each bivariate copula which corresponds to an edge in the graphical structure. If a d-dimensional regular vine copula is meant to be specified then the set of bivariate copulas has $d(d-1)/2$ elements (Czado, 2020). The choice between alternative copula families is made based on the Akaike Information Criterion (AIC), a likelihood-based measure. The parametric

[70] In the case that no two subjects are tied with respect to the observed values of any of the variables, the summation component of the specification is equivalent to the ranking of subject i within the set of observed values for variable j. This specification implies that in the case that some subjects have the same value for a particular variable the maximum ranking possible is assigned to each of those subjects. However, for the implementation in this paper, in case of ties the average ranking is assigned to each of the subjects tied such that the sum of the rankings is equivalent to the case of no ties in the data.

copula family with the minimum AIC value out of a pre-specified set of parametric copula families is chosen. For a sample of size n we have bivariate pseudo-copula data on n observations for each edge[71]. As such, the AIC value for each candidate parametric family is calculated through maximum pseudo-likelihood estimation. The estimation method follows the same procedure as maximum likelihood estimation. The amendment in the name of the estimation method represents the fact that pseudo-copula data is used.

The set of parametric copula families considered as candidates in this paper include the Independence, Gaussian, Student t, Clayton, Gumbel, Frank, and Joe families. Additional copula families included which admit a two-parameter specification include the BB1, BB6, BB7, BB8, Tawn type 1, and Tawn type 2 families[72]. When applicable, rotations of the parametric copula families, as described in subsection A.3, are also included in the set of candidates for each edge[73]. Note that the Independence copula is assigned to a particular edge based on an independence test performed before the selection procedure from the set of parametric copula families which is based on AIC[74]. As soon as the set of bivariate copulas corresponding to the edges of a particular tree is specified, the parameters associated to each copula are estimated. Based on the specified set of parametric bivariate copulas and the estimated parameters for any tree in the sequence, the application of the *h-function* as suggested in subsection B.3 is used to generate the pseudo-copula data used for the bivariate copula selection of the subsequent tree. For a given regular vine tree sequence, this procedure which repeats itself until all the bivariate copulas in all the trees have been specified and all the relevant parameters have been estimated is known as sequential estimation (Czado, 2020).

The approach described above assumes that the regular vine tree sequence is given. To determine a specific structure for the graphical component of a regular vine the approach by Dißmann *et al*. (2013) is used in this paper. The so-called Dißmann's algorithm uses the

[71] For the first tree, the pseudo-copula data is obtained by the application of the empirical distribution function on the original data set. For subsequent trees, the pseudo-copula data is obtained through the recursive application of the *h-function* as presented in subsection B.3.

[72] The model selection and estimation of the regular vine copula is carried out using the statistical software R, and in particular the R package VineCopula.
This is available at https://cran.r-project.org/web/packages/VineCopula/VineCopula.pdf. The set of parametric copula families considered is the one provided by the aforementioned package.

[73] Note that some of the parametric families considered do not strictly satisfy the symmetry criterion proposed in subsection B.2 as part of the definition of a regular vine distribution. However, the fact that counterclockwise rotations of copulas by 90, 180, and 270 degrees are considered where applicable implies that in essence the situation is equivalent to considering only symmetric copulas.

[74] The Independence copula is assigned as long as the p-value of the independence test exceeds 0.05. The independence test is performed as described in Genest and Favre (2007), and as implemented in the R package VineCopula.

aforementioned sequential estimation approach, where an additional step of structure selection for each tree in the sequence is embedded in the procedure before the selection and estimation of bivariate copulas. The structure selected for each individual tree is such that the end product conforms with the definition of a regular vine tree sequence as presented in subsection B.2 (e.g. the proximity condition is satisfied).

Starting from the first tree of the sequence, Dißmann's algorithm prescribes a spanning tree as a structure for the first component of the graphical representation. A spanning tree is a tree on all nodes, as defined in subsection B.2. The choice of the specific spanning tree used is made based on an arbitrary criterion. For each possible spanning tree $G = (N, E)$ a weight is assigned to each element of the set E. The spanning tree chosen $G^* = (N^*, E^*)$ is such that the sum of the weights across all elements of the set E^* is greater than that of any other set E which composes a candidate spanning tree $G = (N, E)$. This suggests that the set of edges assumed to exist are those which carry the highest total weight, and make up a spanning tree structure at the same time. Given the spanning tree, the bivariate copula selection and estimation steps follow, as described in the first part of the current subsection. Given the first tree structure, the specification of bivariate copulas, and the estimates of the relevant parameters, pseudo-copula data can be generated for the subsequent tree through the use of the *h-function* as described in subsection B.3. Based on the pseudo-copula data, the same procedure involving weights assigned to edges of candidate spanning trees follows. However, for trees of the sequence after the first one the set of candidate spanning trees is limited to those spanning trees which satisfy the proximity condition, as prescribed by the regular vine tree sequence definition of subsection B.2. Bivariate copula selection and estimation, as well as pseudo-copula data calculation follow again. This process repeats itself until all $d - 1$ trees in the sequence are assigned a structure and a set of bivariate copulas with the relevant estimated parameters.

The structure selected for each tree depends on the definition of the weight assigned to each edge of the candidate spanning trees. As already seen, the edges in each tree of the regular vine tree sequence capture bivariate associations. By using bivariate pseudo-copula data available for the construction of each tree in the sequence, an empirical statistic can be used as a weight for each edge to capture the strength of the bivariate association represented by the edge. In accordance with Dißmann *et al*. (2013), the absolute value of the empirical version of Kendall's tau is used in this paper as the definition of the weight assigned to each edge. Kendall's tau is a measure of rank correlation. Kendall's tau and its empirical versions are defined in subsection 5.6.

Dißmann's algorithm provides a complete procedure for the selection and estimation of a regular vine copula given pseudo-copula data. It is a greedy procedure in the sense that the attempt to achieve optimality in fitting is performed sequentially within each isolated tree rather than e.g. simultaneous model selection in terms of the entire regular vine tree sequence and the corresponding bivariate copulas. In an ideal world every possible regular vine tree sequence would be tested as a candidate for the graphical structure. However, this is not necessarily an efficient approach courtesy of the significant computational expense that comes with it. Even for a relatively small number of dimensions the amount of possible regular vine tree sequences can be vast.

APPENDIX C

C.1 Summary statistics of variables in the bivariate ordinal regression model

Table C.1: Summary statistics of variables used in the bivariate ordinal regression model.

Variable	Sample mean	Sample standard deviation
Year		
2010	0.389	0.488
2011	0.589	0.492
2012	0.022	0.147
Job Status		
Self-employed	0.083	0.276
Paid employment	0.510	0.500
Unemployed	0.036	0.185
Retired	0.267	0.443
On maternity leave	0.005	0.068
Family care	0.049	0.216
Full-time student	0.025	0.155
Long-term sick or Disabled	0.021	0.144
Government training scheme	0.0001	0.012
Unpaid, family business	0.0007	0.026
Doing something else	0.004	0.062
Health		
Excellent	0.150	0.357
Very good	0.382	0.486
Good	0.317	0.465
Fair	0.121	0.326
Poor	0.030	0.170
Country		
England	0.916	0.278
Wales	0.084	0.278
Marital Status		
Single	0.129	0.336
Married	0.592	0.491
Same-sex civil partnership	0.004	0.060
Separated	0.018	0.132
Divorced	0.080	0.272
Widowed	0.060	0.238
Separated from civil partner	0.0001	0.012
Living as couple	0.116	0.321
Children number	0.477	0.869
Education		
Degree	0.229	0.420
Other higher degree	0.131	0.338
A-level etc	0.201	0.401
GCSE etc	0.215	0.411
Other qualification	0.106	0.307
No qualification	0.118	0.322
Logarithm of income	7.476	0.668

Age	51.513	16.126
Agreeableness	5.654	0.995
Extraversion	4.632	1.315
Openness	4.577	1.269
Neuroticism	3.483	1.428
Conscientiousness	5.557	0.995
Sex		
Male	0.448	0.497
Female	0.552	0.497
Life Satisfaction	5.286	1.432
GHQ	-10.950	5.278
Race		
British, English, Scottish, Welsh, Northern Irish	0.930	0.255
Irish	0.008	0.089
Gypsy or Irish traveller	0.0001	0.012
Any other white background	0.026	0.159
White and black Caribbean	0.002	0.047
White and black African	0.002	0.039
White and Asian	0.002	0.045
Any other mixed background	0.001	0.033
Indian	0.010	0.099
Pakistani	0.004	0.062
Bangladeshi	0.001	0.037
Chinese	0.001	0.035
Any other Asian background	0.004	0.063
Caribbean	0.003	0.056
African	0.003	0.055
Any other black background	0.0004	0.020
Arab	0.001	0.033
Any other ethnic group	0.001	0.031

Notes: Sample consists of 7,317 observations. The GHQ scale is reversed such that a higher value in the set {-36,…,0} represents a higher level of well-being.

APPENDIX D

D.1 Estimated regular vine copula

A series of 11 figures presented below is used to represent the 11 trees which constitute the estimated regular vine copula.

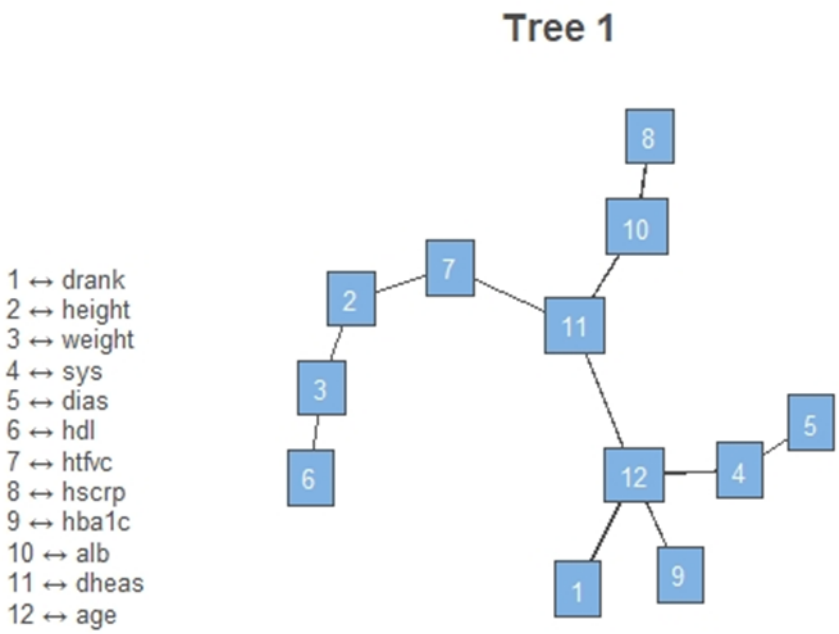

Figure D.1: Tree 1 of the estimated regular vine copula.

The linked variables in *Figure D.1* are the ones for which the unconditional association is modelled. A full list of the chosen copulas along with the estimated parameters for all pairs is provided in *Table D.1*. Recall that a link between two variables is also known as an edge.

Table D.1: Estimated parametric copulas for Tree 1.

Edge	Family	Parameter 1	Parameter 2
3, 6	Frank	-2.77	-
2, 3	Frank	3.60	-
12, 1	BB8	1.32	0.87
7, 2	Student t	0.76	15.03
12, 9	Frank	3.58	-
4, 5	Student t	0.65	11.54
10, 8	Gaussian	-0.25	-
11, 10	Gaussian	0.39	-
11, 7	Gaussian	0.51	-
12, 4	BB8	6.00	0.37
12, 11	Frank	-4.47	-

For the pair of nodes 1 and 12 which represent the well-being variable and the age variable respectively, the BB8 copula family is chosen to model the unconditional relationship between the two. The BB8 copula family is also known as the Joe-Frank copula family. Based on the estimated parameters, the copula does not exhibit tail dependence. Loosely speaking, this means that the strength of the association between the two variables does not change when we consider the extreme values of the two variables[75].

As far as the rest of the pairs are concerned, most of the associations implied by the structure of the first tree are centered around the variables *age* and *dheas*. The associations between *weight* and *hdl*, *hscrp* and *alb*, and *dheas* and *age* are modelled as negative. The rest are modelled as positive associations[76].

The links of the first tree become the nodes of the second tree. In general, the links of any tree become the nodes of the next tree in the sequence of trees which constitutes an estimated regular vine copula. The second tree is presented in *Figure D.2*.

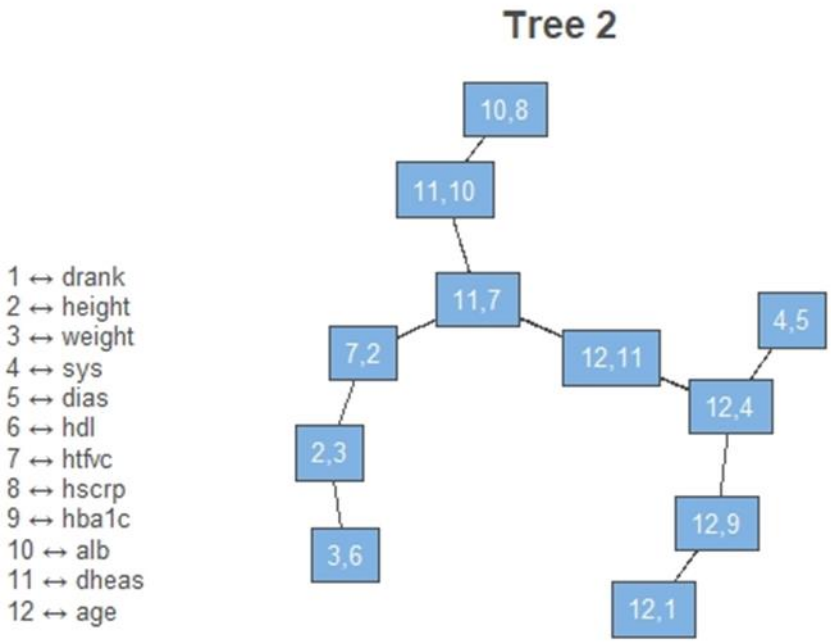

Figure D.2: Tree 2 of the estimated regular vine copula.

[75] The BB8 copula exhibits tail dependence in the case that parameter 2 is equal to 1. This is the case when the BB8 copula family boils down to the Joe copula family.
[76] It should be noted that the estimated parameter 2 for the BB8 copula for the edge between *sys* and *age* is on the boundary of the constraint imposed for a maximum of 6.

In the second tree, a link between two nodes represents the existence of a conditional relationship[77]. The variables which appear in both nodes of the link are the conditioning variables. The remaining two variables are the ones for which the conditional association is modelled. This is a general rule applying to the links of all remaining trees. A list of the chosen copulas along with the estimated parameters is provided in *Table D.2*.

Table D.2: Estimated parametric copulas for Tree 2.

Edge	Family	Parameter 1	Parameter 2
2, 6; 3	Frank	-0.17	-
7, 3; 2	BB8 90^o	-2.44	-0.43
9, 1; 12	Clayton 90^o	-0.09	-
11, 2; 7	Gaussian	-0.05	-
4, 9; 12	BB8	1.15	0.84
12, 5; 4	Tawn type 1 270^o	-1.51	0.38
11, 8; 10	BB8 90^o	-1.31	-0.66
7, 10; 11	Student t	0.21	30.00
12, 7; 11	Clayton 90^o	-0.36	-
11, 4; 12	Frank	1.05	-

Notes: The semicolon separates the variables (left) for which the conditional association is modelled from the conditioning variables (right).

For the pair of nodes 12, 1 and 12, 9 which represent the age and well-being variables, and the age and glycated haemoglobin variables respectively, the Clayton copula family with a counterclockwise rotation of 90^o is chosen to model the bivariate conditional association of the existing link. Based on the chosen copula, there is no evidence of tail dependence.

Most of the conditional associations implied by the structure of the second tree are centered around 11, 7 and 12, 4. Moreover, the conditional associations between *htfvc* and *alb*, and *sys* and *dheas* are modelled as positive. The rest of the conditional associations are modelled as negative[78].

The links of the second tree become the nodes of the third tree and the resulting tree is presented in *Figure D.3*, and the chosen copulas along with the estimated parameters are provided in *Table D.3*.

[77] The conditional relationship is in the form of a conditional bivariate copula for which the arguments are two conditional distribution functions.

[78] It should be noted that the estimated parameter 2 for the Student t copula family of the edge between 11, 7 and 11, 10 is on the boundary of the constraint imposed for a maximum of 30.

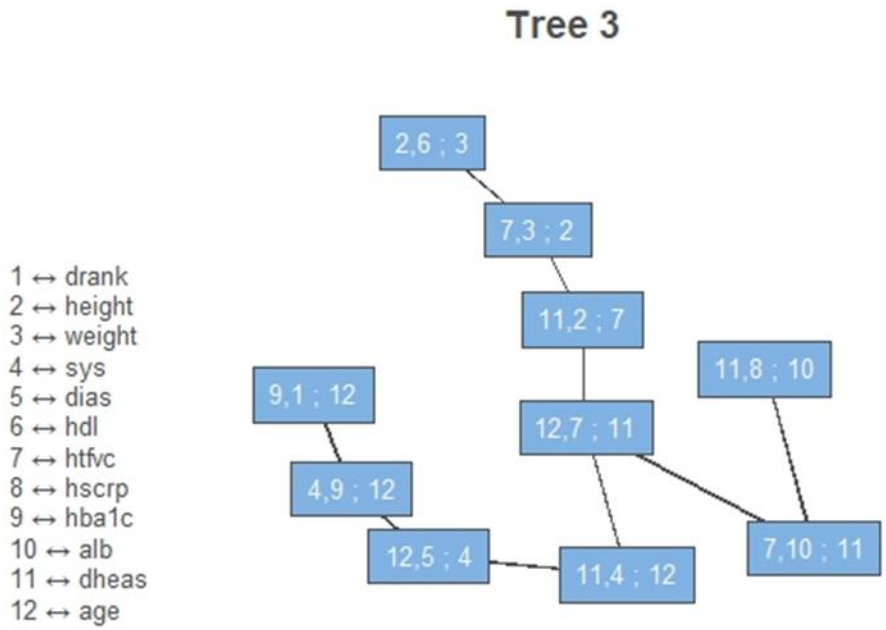

Figure D.3: Tree 3 of the estimated regular vine copula.

Table D.3: Estimated parametric copulas for Tree 3.

Edge	Family	Parameter 1	Parameter 2
7, 6; 2, 3	Frank	-0.29	-
11, 3; 7, 2	Joe 270^{o}	-1.03	-
4, 1; 9, 12	Independence	-	-
12, 2; 11, 7	Frank	1.49	-
5, 9; 4, 12	Independence	-	-
11, 5; 12, 4	BB1 90^{o}	-0.03	-1.03
7, 8; 11, 10	Gaussian	-0.16	-
12, 10; 7, 11	Tawn type 1 270^{o}	-1.21	0.28
4, 7; 12, 11	BB8 180^{o}	1.65	0.56

Notes: The semicolon separates the variables (left) for which the conditional association is modelled from the conditioning variables (right).

For the link between 9, 1; 12 and 4, 9; 12 the Independence copula is chosen to model the conditional association between *drank* and *sys* given *hba1c* and *age*. Based on the chosen copula, well-being and systolic blood pressure are independent given glycated haemoglobin and age [79].

[79] The inference of conditional independence is based on the failure to reject the null of independence at the 5% significance level.

The conditional associations between *height* and *age*, and *sys* and *htfvc* are modelled as positive. The Independence copula is chosen for the conditional association between *dias* and *hba1c*. The rest of the conditional associations are modelled as negative.

The fourth tree is presented in *Figure D.4* and the chosen copulas with the estimated parameters are provided in *Table D.4*.

Table D.4: Estimated parametric copulas for Tree 4.

Edge	Family	Parameter 1	Parameter 2
11, 6; 7, 2, 3	Frank	-0.45	-
12, 3; 11, 7, 2	BB8 180^o	1.17	0.99
5, 1; 4, 9, 12	Gaussian	-0.06	-
4, 2; 12, 11, 7	Frank	0.59	-
11, 9; 5, 4, 12	Independence	-	-
7, 5; 11, 12, 4	Tawn type 2 90^o	-1.09	0.23
12, 8; 7, 11, 10	Tawn type 1 180^o	1.15	0.14
4, 10; 12, 7, 11	Frank	0.72	-

Notes: The semicolon separates the variables (left) for which the conditional association is modelled from the conditioning variables (right).

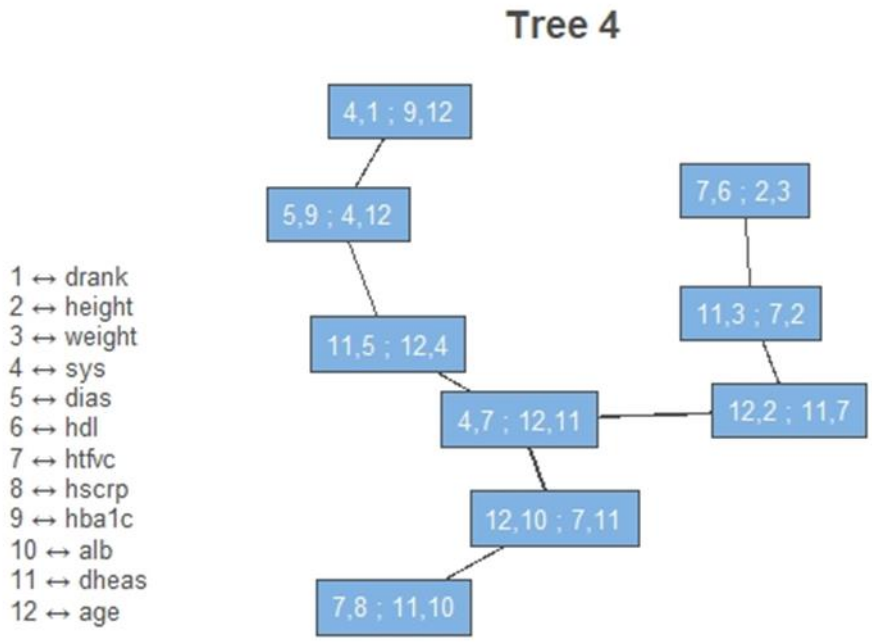

Figure D.4: Tree 4 of the estimated regular vine copula.

The conditional association between well-being and diastolic blood pressure given systolic blood pressure, glycated haemoglobin, and age is represented by the link between 4, 1; 9, 12 and 5, 9; 4, 12, and modelled by the Gaussian copula. There is no evidence of tail dependence.

Evidence of negative conditional associations exists for the pairs *hdl* and *dheas*, and *dias* and *htfvc*. The variables *hba1c* and *dheas* are conditionally independent. The rest of the conditional associations are modelled as positive.

Figure D.5 presents the fifth tree of the sequence, and *Table D.5* the relevant copula choices. In the fifth tree the conditional association between well-being and dehydroepiandrosterone sulphate given diastolic blood pressure, systolic blood pressure, glycated haemoglobin, and age is modelled by the Clayton copula.

The conditional associations between *height* and *dias*, and *htfvc* and *hba1c* are estimated to be negative. The Independence copula is chosen for the conditional association between *dias* and *alb*. The rest of the estimated associations are positive.

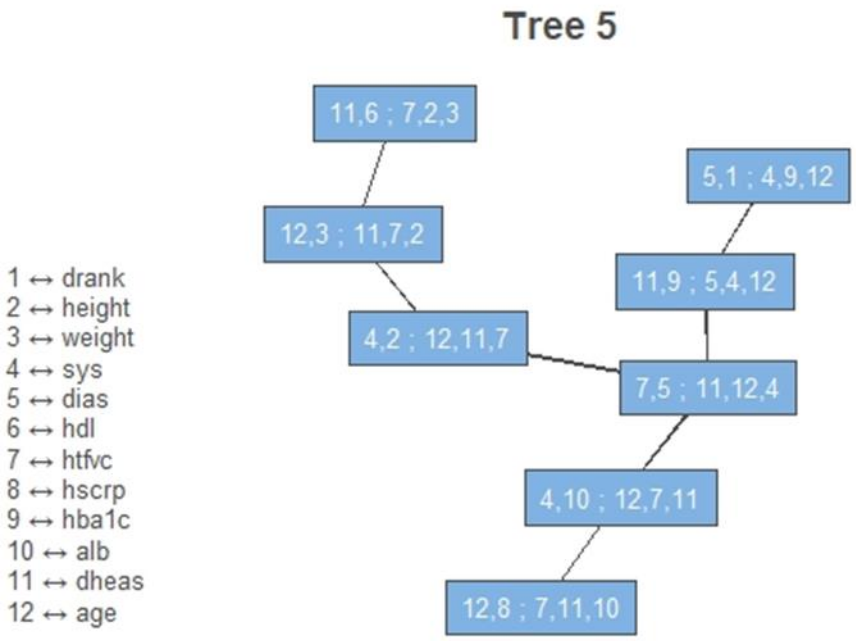

Figure D.5: Tree 5 of the estimated regular vine copula.

Table D.5: Estimated parametric copulas for Tree 5.

Edge	Family	Parameter 1	Parameter 2
12, 6; 11, 7, 2, 3	Tawn type 2 180°	1.15	0.05
4, 3; 12, 11, 7, 2	BB8 180°	2.11	0.65
11, 1; 5, 4, 9, 12	Clayton	0.05	-
5, 2; 4, 12, 11, 7	Gaussian	-0.08	-
7, 9; 11, 5, 4, 12	Clayton 270°	-0.11	-
10, 5; 7, 11, 12, 4	Independence	-	-
4, 8; 12, 7, 11, 10	BB8 180°	2.13	0.53

Notes: The semicolon separates the variables (left) for which the conditional association is modelled from the conditioning variables (right).

The edges of the fifth tree become the nodes of the sixth tree and the resulting tree is presented in *Figure D.6*. A full list of the chosen copulas with the estimated parameters is provided in *Table D.6*.

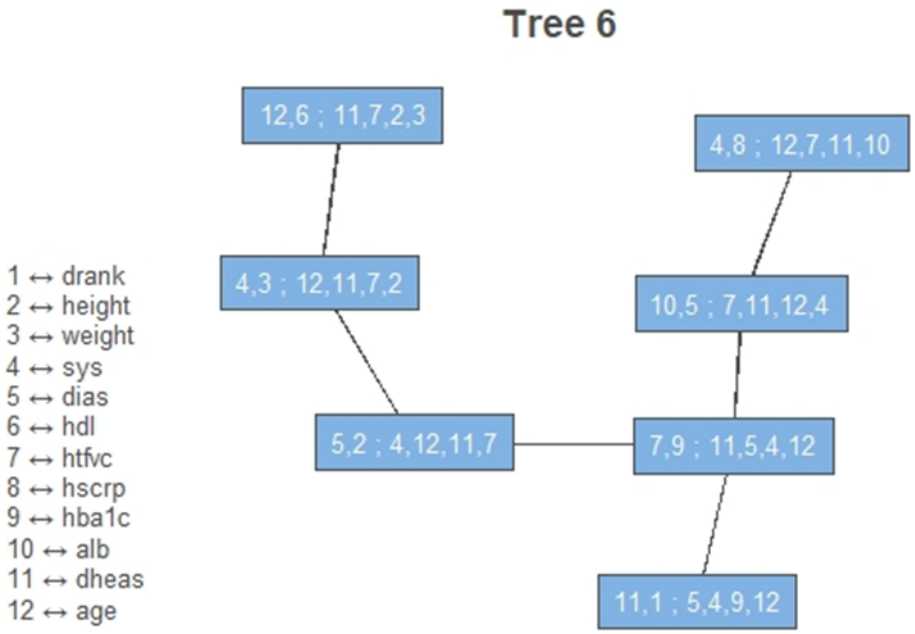

Figure D.6: Tree 6 of the estimated regular vine copula.

Table D.6: Estimated parametric copulas for Tree 6.

Edge	Family	Parameter 1	Parameter 2
4, 6; 12, 11, 7, 2, 3	Independence	-	-
5, 3; 4, 12, 11, 7, 2	BB1 180^o	0.20	1.02
7, 1; 11, 5, 4, 9, 12	Clayton	0.07	-
9, 2; 5, 4, 12, 11, 7	BB8	1.30	0.81
10, 9; 7, 11, 5, 4, 12	Gumbel 90^o	-1.03	-
8, 5; 10, 7, 11, 12, 4	Frank	0.47	-

Notes: The semicolon separates the variables (left) for which the conditional association is modelled from the conditioning variables (right).

The Clayton copula is chosen to model the conditional association between well-being and forced vital capacity given dehydroepiandrosterone sulphate, diastolic blood pressure, systolic blood pressure, glycated haemoglobin, and age.

The conditional association between *hba1c* and *alb* is accompanied by a negative Kendall's tau. The Independence copula is chosen for the conditional association between *sys* and *hdl*. The rest of the estimated conditional associations are positive.

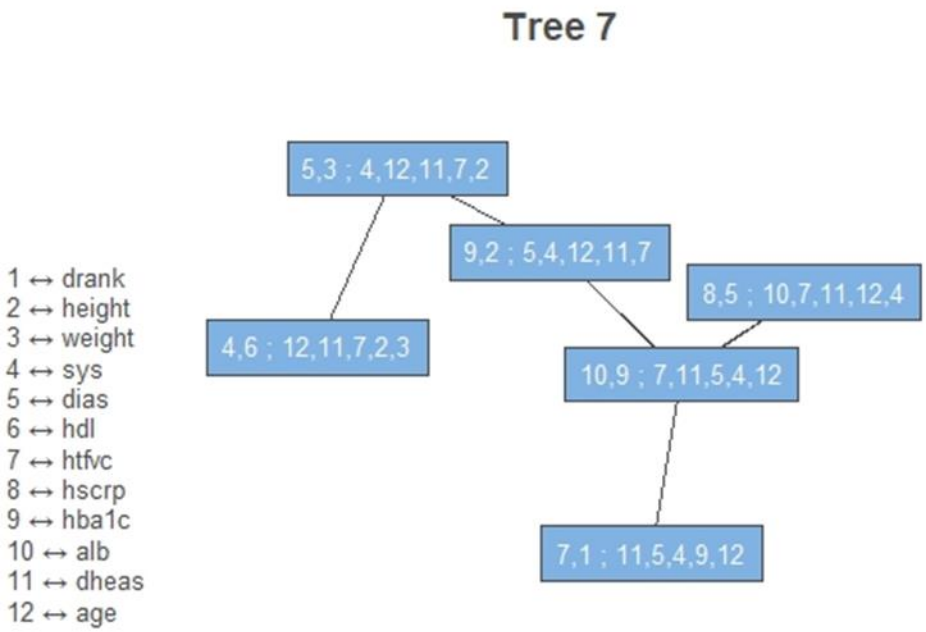

Figure D.7: Tree 7 of the estimated regular vine copula.

Table D.7: Estimated parametric copulas for Tree 7.

Edge	Family	Parameter 1	Parameter 2
5, 6; 4, 12, 11, 7, 2, 3	Gaussian	0.06	-
9, 3; 5, 4, 12, 11, 7, 2	BB8	1.37	0.91
10, 1; 7, 11, 5, 4, 9, 12	BB8 180^o	1.23	0.66
10, 2; 9, 5, 4, 12, 11, 7	Independence	-	-
8, 9; 10, 7, 11, 5, 4, 12	Gaussian	0.15	-

Notes: The semicolon separates the variables (left) for which the conditional association is modelled from the conditioning variables (right).

The seventh tree is presented in *Figure D.7* along with the relevant copulas in *Table D.7*. The BB8 copula family with a rotation of 180^o is chosen as the appropriate one for the conditional association between well-being and albumin given forced vital capacity, dehydroepiandrosterone sulphate, diastolic blood pressure, systolic blood pressure, glycated haemoglobin, and age. The estimated copula does not exhibit tail dependence.

The Independence copula is chosen for the conditional association between *height* and *alb*. The rest of the estimated conditional associations are positive.

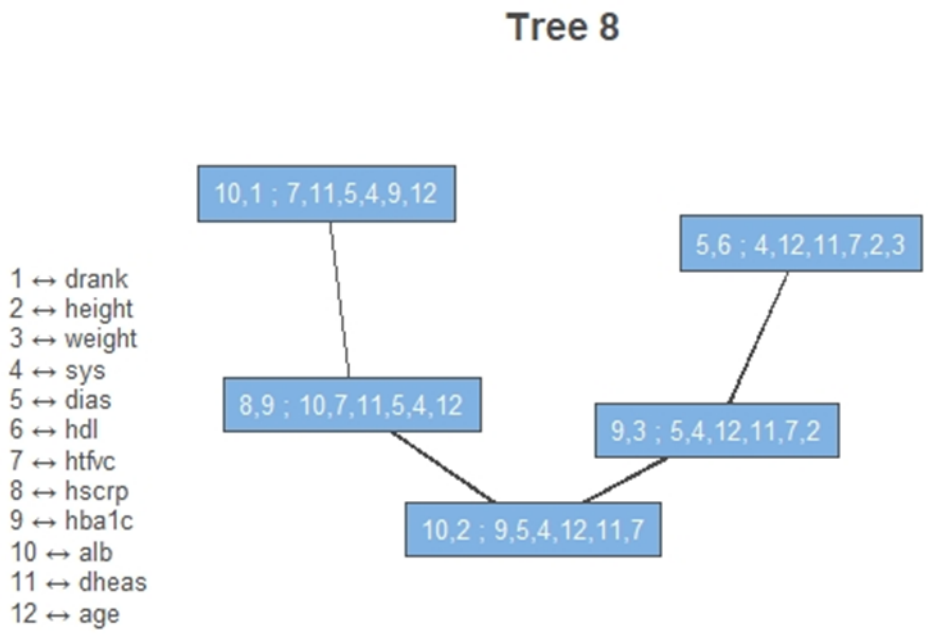

Figure D.8: Tree 8 of the estimated regular vine copula.

Table D.8: Estimated parametric copulas for Tree 8.

Edge	Family	Parameter 1	Parameter 2
9, 6; 5, 4, 12, 11, 7, 2, 3	BB1 270^o	-0.03	-1.09
10, 3; 9, 5, 4, 12, 11, 7, 2	Tawn type 2 90^o	-1.18	0.19
8, 1; 10, 7, 11, 5, 4, 9, 12	Independence	-	-
8, 2; 10, 9, 5, 4, 12, 11, 7	Frank	0.33	-

Notes: The semicolon separates the variables (left) for which the conditional association is modelled from the conditioning variables (right).

The eighth tree is given in *Figure D.8* accompanied by the chosen copulas in *Table D.8*. The Independence copula is the best fit for the conditional association between well-being and c-reactive protein given albumin, forced vital capacity, dehydroepiandrosterone sulphate, diastolic blood pressure, systolic blood pressure, glycated haemoglobin, and age. As such there is no evidence of a significant conditional association between well-being and this particular biomarker.

The conditional association between *height* and *hscrp* is modelled as positive. The rest of the estimated conditional associations are negative.

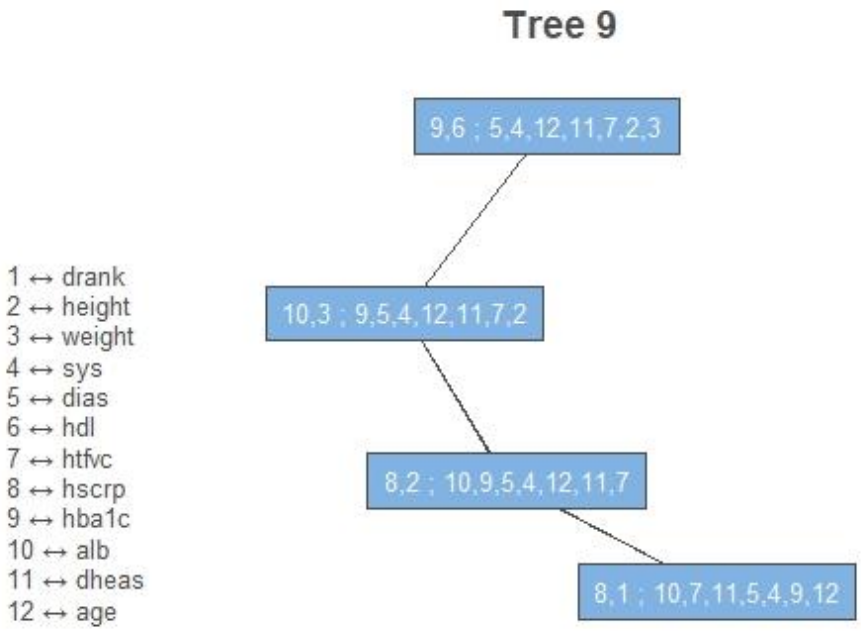

Figure D.9: Tree 9 of the estimated regular vine copula.

Table D.9: Estimated parametric copulas for Tree 9.

Edge	Family	Parameter 1	Parameter 2
10, 6; 9, 5, 4, 12, 11, 7, 2, 3	BB1	0.05	1.05
8, 3; 10, 9, 5, 4, 12, 11, 7, 2	BB8 180^{o}	3.70	0.48
2, 1; 8, 10, 7, 11, 5, 4, 9, 12	Independence	-	-

Notes: The semicolon separates the variables (left) for which the conditional association is modelled from the conditioning variables (right).

Table D.9 provides the chosen copulas for the ninth three presented in *Figure D.9*. The Independence copula is chosen for the conditional association between well-being and height given the variables for c-reactive protein, albumin, forced vital capacity, dehydroepiandrosterone sulphate, diastolic blood pressure, systolic blood pressure, glycated haemoglobin, and age. This implies no significant conditional association between *drank* and *height*. The rest of the conditional associations are modelled as positive.

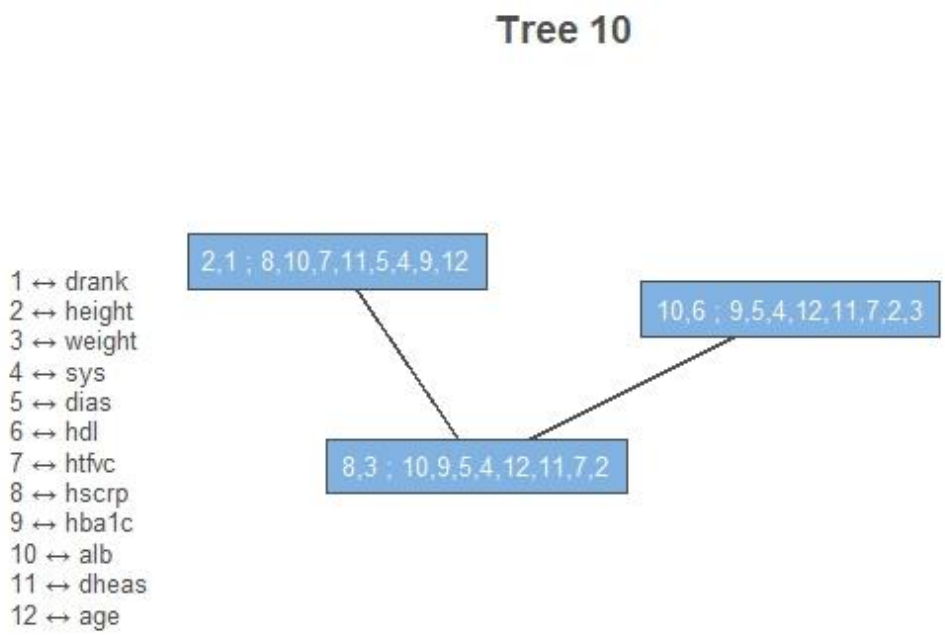

Figure D.10: Tree 10 of the estimated regular vine copula.

Table D.10: Estimated parametric copulas for Tree 10.

Edge	Family	Parameter 1	Parameter 2
8, 6; 10, 9, 5, 4, 12, 11, 7, 2, 3	Student t	-0.08	30.00
1, 3; 8, 10, 9, 5, 4, 12, 11, 7, 2	Independence	-	-

Notes: The semicolon separates the variables (left) for which the conditional association is modelled from the conditioning variables (right).

As it can be seen in *Table D.10* and *Figure D.10*, the Independence copula is chosen in the tenth tree to model the conditional association between well-being and weight given height, c-

reactive protein, albumin, forced vital capacity, dehydroepiandrosterone sulphate, diastolic blood pressure, systolic blood pressure, glycated haemoglobin, and age. As such, no significant conditional association between the two can be inferred. The conditional association between *hdl* and *hscrp* is modelled as negative[80].

Table D.11: Estimated parametric copula for Tree 11.

Edge	Family	Parameter 1	Parameter 2
1, 6; 8, 10, 9, 5, 4, 12, 11, 7, 2, 3	Clayton	0.03	-

Notes: The semicolon separates the variables (left) for which the conditional association is modelled from the conditioning variables (right).

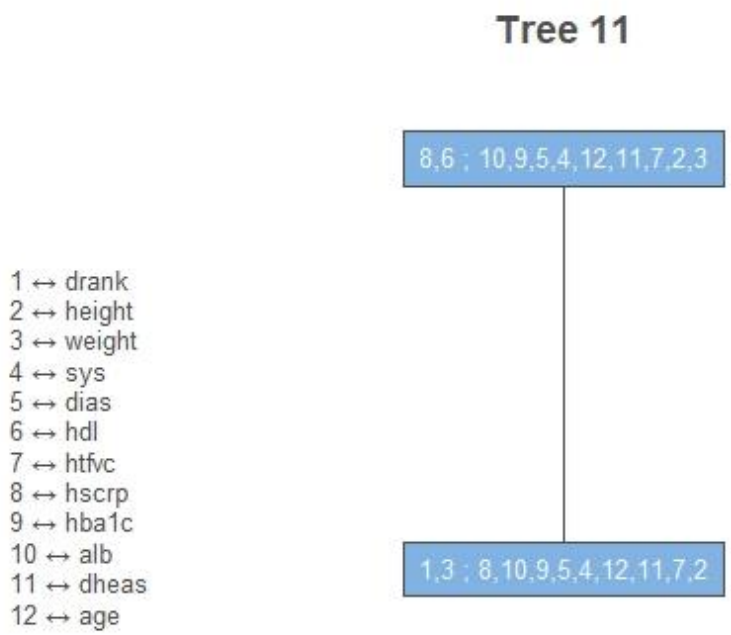

Figure D.11: Tree 11 of the estimated regular vine copula.

From *Table D.11* and *Figure D.11*, the Clayton copula is chosen to model the conditional association between well-being and high-density lipoprotein given weight, height, c-reactive protein, albumin, forced vital capacity, dehydroepiandrosterone sulphate, diastolic blood pressure, systolic blood pressure, glycated haemoglobin, and age.

[80] It should be noted that the estimated parameter 2 for the Student t copula family is on the boundary of the constraint imposed for a maximum of 30.

D.2 Robustness checks

Table D.12: Estimated regular vine copula based on random ranking with number seed 111.

Tree	Edge	Family	Parameter 1	Parameter 2
1	3, 6	Frank	-2.76	-
	3, 8	Gaussian	0.25	-
	2, 3	Frank	3.60	-
	12, 1	BB8	1.32	0.87
	7, 2	Student t	0.76	15.02
	12, 9	Frank	3.55	-
	4, 5	Student t	0.65	11.52
	11, 10	Gaussian	0.38	-
	11, 7	Gaussian	0.51	-
	12, 4	BB8	6.00	0.37
	12, 11	Frank	-4.47	-
2	8, 6; 3	Gaussian	-0.08	-
	2, 8; 3	Frank	-2.02	-
	7, 3; 2	BB8 90^o	-2.45	-0.42
	9, 1; 12	Clayton 90^o	-0.09	-
	11, 2; 7	Gaussian	-0.05	-
	4, 9; 12	BB8	1.14	0.85
	12, 5; 4	Tawn type 1 270^o	-1.51	0.38
	7, 10; 11	Student t	0.21	30.00
	12, 7; 11	Clayton 90^o	-0.36	-
	11, 4; 12	Frank	1.05	-
3	2, 6: 8, 3	Frank	-0.33	
	7, 8: 2, 3	Student t	-0.18	30.00
	11, 3; 7, 2	Joe 270^o	-1.03	-
	4, 1; 9, 12	Independence	-	-
	12, 2; 11, 7	Frank	1.49	-
	5, 9; 4, 12	Independence	-	-
	11, 5; 12, 4	BB1 90^o	-0.03	-1.03
	12, 10; 7, 11	Tawn type 1 270^o	-1.21	0.28
	4, 7; 12, 11	BB8 180^o	1.64	0.56
4	7, 6; 2, 8, 3	Frank	-0.41	-
	11, 8; 7, 2, 3	Clayton 270^o	-0.04	-
	12, 3; 11, 7, 2	BB8 180^o	1.17	0.99
	5, 1; 4, 9, 12	Gaussian	-0.06	-
	4, 2; 12, 11, 7	Frank	0.59	-
	11, 9; 5, 4, 12	Independence	-	-
	7, 5; 11, 12, 4	Tawn type 2 90^o	-1.09	0.23
	4, 10; 12, 7, 11	Frank	0.71	-
5	11, 6; 7, 2, 8, 3	Frank	-0.47	-
	12, 8; 11, 7, 2, 3	Independence	-	-
	4, 3; 12, 11, 7, 2	BB8 180^o	2.10	0.65
	11, 1; 5, 4, 9, 12	Clayton	0.04	-
	5, 2; 4, 12, 11, 7	Gaussian	-0.08	-
	7, 9; 11, 5, 4, 12	Clayton 270^o	-0.11	-
	10, 5; 7, 11, 12, 4	Independence	-	-
6	12, 6; 11, 7, 2, 8, 3	Tawn type 2 180^o	1.16	0.04

	4, 8; 12, 11, 7, 2, 3	Gaussian	0.05	-
	5, 3; 4, 12, 11, 7, 2	BB1 180º	0.20	1.02
	7, 1; 11, 5, 4, 9, 12	Clayton	0.08	-
	9, 2; 5, 4, 12, 11, 7	BB8	1.29	0.82
	10, 9; 7, 11, 5, 4, 12	Gumbel 90º	-1.03	-
7	4, 6; 12, 11, 7, 2, 8, 3	Independence	-	-
	5, 8; 4, 12, 11, 7, 2, 3	Independence	-	-
	9, 3; 5, 4, 12, 11, 7, 2	BB8	1.37	0.91
	10, 1; 7, 11, 5, 4, 9, 12	Frank	0.30	-
	10, 2; 9, 5, 4, 12, 11, 7	Independence	-	-
8	5, 6; 4, 12, 11, 7, 2, 8, 3	Gaussian	0.06	-
	9, 8; 5, 4, 12, 11, 7, 2, 3	Gaussian	0.09	-
	10, 3; 9, 5, 4, 12, 11, 7, 2	Tawn type 2 90º	-1.17	0.20
	2, 1; 10, 7, 11, 5, 4, 9, 12	Independence	-	-
9	9, 6; 5, 4, 12, 11, 7, 2, 8, 3	BB7 270º	-1.10	-0.07
	10, 8; 9, 5, 4, 12, 11, 7, 2, 3	BB8 90º	-1.48	-0.77
	1, 3; 10, 9, 5, 4, 12, 11, 7, 2	Independence	-	-
10	10, 6; 9, 5, 4, 12, 11, 7, 2, 8, 3	BB1	0.04	1.04
	1, 8; 10, 9, 5, 4, 12, 11, 7, 2, 3	Independence	-	-
11	1, 6; 10, 9, 5, 4, 12, 11, 7, 2, 8, 3	Clayton	0.03	-

Notes: The number coding is identical to the one used in Figures D.1-D.11 in the current section. The semicolon separates the variables (left) for which the conditional association is modelled from the conditioning variables (right).

The estimated regular vine copula which corresponds to the incorporation of random ranking is presented in *Table D.12*. This is the estimation which uses the number seed 111. Due to space considerations and the fact that only minor differences exist between them, the rest of the estimated regular vine copulas which correspond to the remainder of the number seeds used are not included but are available on request[81]. The differences between them which are of interest to this paper are reported in the current subsection. In addition, *Table D.13* provides the estimated regular vine copula which uses the well-being measure of subsection 5.6[82].

Starting from the repetitive process which involves the pseudo-random ranking generation, there appears to be stability in the (conditional) associations which involve the well-being variable. This stability is defined in terms of the order in which these associations are modelled in the relevant trees, the direction of association, and the bivariate copula selection for each association. To be more precise, based on the three aforementioned criteria of stability, the initial six trees of each estimated tree sequence appear to be identical to the original estimation

[81] The term 'minor' in this case is used to indicate the fact that the order in which bivariate (conditional) associations are modelled is identical across the 10 repetitions, as well as the direction of association in terms of being positive or negative. What differs between some of the repetitions is the choice of the appropriate bivariate copula used to model the association for few pairs of variables in each estimated regular vine copula.

[82] Note that there is a reduced sample size of 7,317 due to the way in which the well-being variable is constructed. It should also be noted that in this estimation, the method based on which ties are dealt with is the same as in the original estimation.

when it comes to modelling the well-being variable. Minor differences are evident in the estimated parameters of the selected bivariate copulas.

The first main difference appears to be in the case of the seventh tree of the sequence. In the original tree, the bivariate conditional association between *drank* and *alb* is modelled using the BB8 copula family with a rotation of 180^{o}. In five of the repetitions, the Frank copula family is used to model this association. However, it should be noted that based on the estimated parameters, the Kendall's tau which corresponds to the association between the two variables is given by 0.03 in all five cases. This is the same value as in the original estimation. Furthermore, for both the Frank family and the BB8 family, the copulas do not exhibit tail dependence. The main difference between the two families is that the Frank family exhibits symmetric dependence in the tails of the distribution whereas the BB8 family does not, but otherwise the two are not very different in terms of their properties.

The second major difference comes in the form of the order in which bivariate conditional associations involving well-being are modelled in trees 8, 9, and 10. In the original estimation, the conditional association between *drank* and *hscrp* is modelled first in tree 8, the one between *drank* and *height* follows, and the one between *drank* and *weight* is modelled in tree 10. In each of the ten repetitions, the one between *drank* and *height* is first, followed by the one between *drank* and *weight*, and lastly the one between *drank* and *hscrp*. However, in all possible cases, including the original estimation, the Independence copula is chosen as the appropriate bivariate copula for each one of the three modelled conditional associations. The last tree of the sequence for each of the ten repetitions appears to be identical to the original estimation when it comes to modelling the well-being variable. Again, minor differences are evident in the estimated parameters.

When it comes to the rest of the variables in the set of interest, the first thing to note is that the direction of association between the different variables does not vary across the 10 estimations which incorporate random ranking when compared to the original estimation. This is true regardless of the order in which the bivariate associations are modelled, or the bivariate copula chosen as the appropriate to model a relationship. This statement excludes the cases when the null hypothesis of independence is rejected, or not rejected, as opposed to the original estimation. The only case when the direction of association appears to change is for the bivariate conditional association between *height* and *hscrp* in the eighth tree of the original estimation. This conditional association is modelled in the second tree for each of the 10

random ranking repetitions. In the original estimation, the conditional bivariate association between *height* and *hscrp* is modelled as positive, whereas in each of the repetitions it is modelled as negative[83].

The rest of the differences between the original estimation and each of the 10 estimated models using random ranking have to do with the order in which bivariate associations are modelled, as well as with the selected copula families appropriate for each association. Even though the differences are few, they will not be examined in this subsection as the focus lies with the examination of well-being associations.

[83] Note that the association in the original estimation is conditional on *alb*, *hba1c*, *dias*, *sys*, *age*, *dheas*, and *htfvc*. In each of the repetitions the association is conditional on *weight*.

Table D.13: Estimated regular vine copula model based on alternative well-being variable.

Tree	Edge	Family	Parameter 1	Parameter 2
1	3, 6	Frank	-2.76	-
	2, 3	Frank	3.58	-
	12, 1	BB8	1.48	0.89
	7, 2	Student t	0.76	15.32
	12, 9	BB8 180^o	6.00	0.48
	4, 5	Student t	0.65	11.27
	10, 8	Gaussian	-0.25	-
	11, 10	Gaussian	0.38	-
	11, 7	Gaussian	0.51	-
	12, 4	BB8	6.00	0.36
	12, 11	Frank	-4.34	-
2	2, 6; 3	Frank	-0.15	-
	7, 3; 2	BB8 90^o	-2.09	-0.51
	9, 1; 12	Clayton 90^o	-0.10	-
	11, 2; 7	Frank	-0.28	-
	4, 9; 12	BB8	1.17	0.82
	12, 5; 4	Tawn type 1 270^o	-1.50	0.38
	11, 8; 10	BB8 90^o	-1.20	-0.78
	7, 10; 11	Gaussian	0.21	-
	12, 7; 11	Clayton 90^o	-0.36	-
	11, 4; 12	Frank	1.06	-
3	7, 6; 2, 3	Frank	-0.27	-
	11, 3; 7, 2	Independence	-	-
	4, 1; 9, 12	Student t	-0.02	24.09
	12, 2; 11, 7	Frank	1.50	-
	5, 9; 4, 12	BB8 90^o	-1.06	-0.96
	11, 5; 12, 4	BB1 90^o	-0.03	-1.03
	7, 8; 11, 10	Gaussian	-0.16	-
	12, 10; 7, 11	Tawn type 1 270^o	-1.21	0.27
	4, 7; 12, 11	Tawn type 2 180^o	1.20	0.16
4	11, 6; 7, 2, 3	Frank	-0.45	-
	12, 3; 11, 7, 2	BB8 180^o	1.16	1.00
	5, 1; 4, 9, 12	Gaussian	-0.06	-
	4, 2; 12, 11, 7	Frank	0.57	-
	11, 9; 5, 4, 12	Independence	-	-
	7, 5; 11, 12, 4	Tawn type 2 90^o	-1.10	0.22
	12, 8; 7, 11, 10	Tawn type 1 180^o	1.16	0.11
	4, 10; 12, 7, 11	Frank	0.69	-
5	12, 6; 11, 7, 2, 3	Tawn type 2 180^o	1.20	0.04
	4, 3; 12, 11, 7, 2	BB8 180^o	2.14	0.65
	11, 1; 5, 4, 9, 12	Independence	-	-
	5, 2; 4, 12, 11, 7	Gaussian	-0.08	-
	7, 9; 11, 5, 4, 12	Clayton 270^o	-0.10	-
	10, 5; 7, 11, 12, 4	Independence	-	-
	4, 8; 12, 7, 11, 10	BB8 180^o	2.21	0.52
6	4, 6; 12, 11, 7, 2, 3	Independence	-	-
	5, 3; 4, 12, 11, 7, 2	BB1 180^o	0.19	1.02

	7, 1; 11, 5, 4, 9, 12	Clayton	0.06	-
	9, 2; 5, 4, 12, 11, 7	BB8	1.35	0.80
	10, 9; 7, 11, 5, 4, 12	Gumbel 90^{o}	-1.03	-
	8, 5; 10, 7, 11, 12, 4	BB8	1.55	0.53
7	5, 6; 4, 12, 11, 7, 2, 3	Tawn type 1	1.10	0.14
	9, 3; 5, 4, 12, 11, 7, 2	BB8	1.38	0.90
	10, 1; 7, 11, 5, 4, 9, 12	Frank	0.29	-
	10, 2; 9, 5, 4, 12, 11, 7	Independence	-	-
	8, 9; 10, 7, 11, 5, 4, 12	Gaussian	0.15	-
8	9, 6; 5, 4, 12, 11, 7, 2, 3	BB1 270^{o}	-0.02	-1.09
	10, 3; 9, 5, 4, 12, 11, 7, 2	Tawn type 2 90^{o}	-1.21	0.15
	8, 1; 10, 7, 11, 5, 4, 9, 12	Independence	-	-
	8, 2; 10, 9, 5, 4, 12, 11, 7	Frank	0.27	-
9	10, 6; 9, 5, 4, 12, 11, 7, 2, 3	BB1	0.06	1.05
	8, 3; 10, 9, 5, 4, 12, 11, 7, 2	BB8 180^{o}	4.35	0.41
	2, 1; 8, 10, 7, 11, 5, 4, 9, 12	Independence	-	-
10	8, 6; 10, 9, 5, 4, 12, 11, 7, 2, 3	Student t	-0.08	30.00
	1, 3; 8, 10, 9, 5, 4, 12, 11, 7, 2	Clayton 270^{o}	-0.04	-
11	1, 6; 8, 10, 9, 5, 4, 12, 11, 7, 2, 3	BB8 180^{o}	1.20	0.71

Notes: The number coding is identical to the one used in Figures D.1-D.11 in the current section. The semicolon separates the variables (left) for which the conditional association is modelled from the conditioning variables (right).

Moving on to the estimation which uses the constructed measure of subsection 5.6, the initial two trees of the estimated sequence appear to be identical to the original estimation when it comes to modelling the well-being variable. Minor differences are evident in the estimated parameters of the selected bivariate copulas. The first main difference appears to be in the case of the third tree of the sequence. In the original tree, the bivariate conditional association between *drank* and *sys* is modelled using the Independence copula. In the estimation using *rank*[84], the Student t copula family is used to model this association, indicating a negative conditional relationship between the two. Based on the estimated parameters, the Kendall's tau which corresponds to the association between the two variables is given by -0.02. This implies that in the case in which *rank* is used instead of *drank*, for the independence test performed, the null hypothesis of independence is rejected.

The fourth tree of the sequence appears to be identical to the original estimation when it comes to modelling the well-being variable. The second main difference appears to be in the case of the fifth tree of the sequence. In the original tree, the bivariate conditional association between *drank* and *dheas* is modelled using the Clayton copula family. In the estimation using *rank*, the Independence copula is used to model this association. This implies that in the case in which

[84] This term is used to indicate the measure of subsection 5.6.

rank is used instead of *drank*, for the independence test performed, the null hypothesis of independence cannot be rejected.

The sixth tree of the sequence appears to be identical to the original estimation when it comes to modelling the well-being variable, apart from a minor difference in the estimated parameter. The third main difference appears to be in the case of the seventh tree of the sequence, and it is similar to the difference in some of the repetitions performed using random ranking. In particular, the Frank copula family is used to model the bivariate conditional association between *drank* and *alb* instead of the BB8 copula family with a rotation of 180^{o}. Again, based on the estimated parameters, the Kendall's tau which corresponds to the association between the two variables is given by 0.03.

The eighth and ninth trees of the sequence appear to be identical to the original estimation when it comes to modelling the well-being variable. The fourth main difference appears to be in the case of the tenth tree of the sequence. In the original tree, the bivariate conditional association between *drank* and *weight* is modelled using the Independence copula. In the estimation using *rank*, the Clayton copula family with a rotation of 270^{o} is used to model this association, indicating a negative conditional relationship between the two. Based on the estimated parameters, the Kendall's tau which corresponds to the association between the two variables is given by -0.02. This implies that in the case in which *rank* is used instead of *drank*, for the independence test performed, the null hypothesis of independence is rejected.

The last main difference appears to be in the case of the last tree of the sequence. In the original tree, the bivariate conditional association between *drank* and *hdl* is modelled using the Clayton copula family. In the estimation using *rank*, the BB8 copula family with a rotation of 180^{o} is used to model this association, indicating a positive conditional relationship between the two. However, it should be noted that based on the estimated parameters, the Kendall's tau which corresponds to the association between the two variables is given by 0.03. The value in the original estimation is 0.02. The main difference is that the Clayton copula family exhibits lower tail dependence whereas the BB8 family does not. However, based on the estimated parameters, the coefficient of lower tail dependence in the case of the original estimation is given by 0.00 when approximated to 2 decimal places.

As a last point in this subsection, a comparison can be made between the original estimation with respect to the modelled associations outside those which incorporate well-being. The estimations differ in terms of the well-being measures, but other than that, the rest of the

variables are treated in the same way. As such, it is expected that not many differences exist between the two estimations when it comes to the inference on variables outside the well-being measure. As expected, the structure of both estimated regular vine copulas is identical in terms of the order in which bivariate (conditional) associations are modelled. In addition, the direction of association between the different variables appears to be identical for the two estimations, apart from few cases when the null hypothesis of independence is rejected, or not rejected, as opposed to the original estimation. In terms of the copula families chosen to model the different (conditional) relationships, few differences exist between the two estimations outside those which include the well-being variable mentioned in this subsection. Again, these differences will not be examined in detail in this subsection as the focus lies elsewhere.

Overall, it can be seen that the differences between the original estimation and each of the estimated models used as a robustness check are not prohibitive in terms of drawing inferences for the joint distribution of the variables. This is especially true in the case of the well-being variable for which the differences are outlined in detail in the current subsection.

APPENDIX E

E.1 Estimated regular vines by gender

Table E.1: Estimated regular vine copula based on males only.

Tree	Edge	Family	Parameter 1	Parameter 2
1	12, 1	BB8	1.38	0.86
	12, 9	BB8 180^o	4.87	0.54
	3, 6	Gaussian	-0.31	-
	5, 4	Tawn type 2	2.19	0.66
	3, 5	BB8 180^o	3.39	0.44
	2, 3	Student t	0.41	27.11
	7, 2	Student t	0.61	17.35
	7, 8	Student t	-0.28	30.00
	12, 7	BB8 270^o	-4.02	-0.74
	12, 10	Student t	-0.50	16.32
	12, 11	Frank	-5.80	-
2	7, 1; 12	Clayton	0.07	-
	7, 9; 12	BB8 90^o	-1.46	-0.86
	2, 6; 3	Gaussian	0.12	-
	3, 4; 5	Independence	-	-
	2, 5; 3	Tawn type 1 270^o	-1.23	0.16
	7, 3; 2	Frank	-1.00	-
	8, 2; 7	Frank	0.40	-
	12, 8; 7	Clayton	0.13	-
	11, 7; 12	Clayton	0.11	-
	11, 10; 12	BB7	1.02	0.14
3	9, 1; 7, 12	Clayton 90^o	-0.06	-
	8, 9; 7, 12	Student t	0.13	24.35
	5, 6; 2, 3	Student t	0.07	26.86
	2, 4; 3, 5	Tawn type 2 90^o	-1.31	0.14
	7, 5; 2, 3	Clayton	0.07	-
	8, 3; 7, 2	Tawn type 2 180^o	1.42	0.39
	12, 2; 8, 7	Independence	-	-
	11, 8; 12, 7	Independence	-	-
	10, 7; 11, 12	Gumbel 180^o	1.04	-
4	8, 1; 9, 7, 12	Independence	-	-
	2, 9; 8, 7, 12	Clayton 180^o	0.07	-
	4, 6; 5, 2, 3	BB7	1.02	0.08
	7, 4; 2, 3, 5	BB8 90^o	-1.34	-0.91
	8, 5; 7, 2, 3	Frank	0.38	-
	12, 3; 8, 7, 2	Independence	-	-
	11, 2; 12, 8, 7	Gaussian	-0.04	-
	10, 8; 11, 12, 7	Gaussian	-0.14	-
5	2, 1; 8, 9, 7, 12	Independence	-	-
	3, 9; 2, 8, 7, 12	BB8	1.26	0.94
	7, 6; 4, 5, 2, 3	Independence	-	-
	8, 4; 7, 2, 3, 5	Independence	-	-
	12, 5; 8, 7, 2, 3	Independence	-	-
	11, 3; 12, 8, 7, 2	Student t	-0.03	15.78

	10, 2; 11, 12, 8, 7	Independence	-	-
6	3, 1; 2, 8, 9, 7, 12	Independence	-	-
	5, 9; 3, 2, 8, 7, 12	BB8 90^{o}	-1.40	-0.64
	8, 6; 7, 4, 5, 2, 3	Student t	-0.10	24.50
	12, 4; 8, 7, 2, 3, 5	BB8	1.31	0.95
	11, 5; 12, 8, 7, 2, 3	Frank	0.34	-
	10, 3; 11, 12, 8, 7, 2	Independence	-	-
7	5, 1; 3, 2, 8, 9, 7, 12	BB7 90^{o}	-1.02	-0.05
	4, 9; 5, 3, 2, 8, 7, 12	Independence	-	-
	12, 6; 8, 7, 4, 5, 2, 3	Gaussian	0.13	-
	11, 4; 12, 8, 7, 2, 3, 5	Student t	0.09	23.66
	10, 5; 11, 12, 8, 7, 2, 3	Gaussian	0.13	-
8	4, 1; 5, 3, 2, 8, 9, 7, 12	Student t	0.03	24.35
	6, 9; 4, 5, 3, 2, 8, 7, 12	BB8 90^{o}	-1.12	-1.00
	11, 6; 12, 8, 7, 4, 5, 2, 3	Frank	0.50	-
	10, 4; 11, 12, 8, 7, 2, 3, 5	Gumbel 180^{o}	1.02	-
9	6, 1; 4, 5, 3, 2, 8, 9, 7, 12	Independence	-	-
	11, 9; 6, 4, 5, 3, 2, 8, 7, 12	Independence	-	-
	10, 6; 11, 12, 8, 7, 4, 5, 2, 3	Gaussian	0.13	-
10	11, 1; 6, 4, 5, 3, 2, 8, 9, 7, 12	Independence	-	-
	10, 9; 11, 6, 4, 5, 3, 2, 8, 7, 12	Independence	-	-
11	10, 1; 11, 6, 4, 5, 3, 2, 8, 9, 7, 12	Independence	-	-

Notes: The number coding is identical to the one used in Figures D.1-D.11 in Appendix D. The semicolon separates the variables (left) for which the conditional association is modelled from the conditioning variables (right).

Table E.2: Estimated regular vine copula based on females only.

Tree	Edge	Family	Parameter 1	Parameter 2
1	3, 6	BB8 270^o	-2.39	-0.71
	3, 8	BB8	5.38	0.40
	12, 1	BB8	1.27	0.88
	2, 3	Gaussian	0.30	-
	4, 5	Student t	0.66	9.29
	7, 2	Student t	0.61	18.24
	11, 10	Gaussian	0.27	-
	12, 9	Frank	3.87	-
	12, 4	BB8	5.11	0.53
	12, 7	BB8 270^o	-3.53	-0.84
	12, 11	Frank	-4.40	-
2	8, 6; 3	BB8 270^o	-1.73	-0.64
	2, 8; 3	Student t	-0.26	30.00
	7, 1; 12	BB8 180^o	1.20	0.90
	7, 3; 2	Frank	-0.99	-
	12, 5; 4	BB1 90^o	-0.37	-1.06
	12, 2; 7	Gumbel 180^o	1.05	-
	12, 10; 11	Tawn type 1 270^o	-1.22	0.27
	7, 9; 12	BB8 90^o	-1.23	-0.97
	7, 4; 12	Tawn type 2 90^o	-1.11	0.24
	11, 7; 12	Clayton	0.07	-
3	2, 6; 8, 3	Gaussian	0.10	-
	7, 8; 2, 3	Gumbel 270^o	-1.11	-
	9, 1; 7, 12	Clayton 90^o	-0.08	-
	12, 3; 7, 2	Independence	-	-
	7, 5; 12, 4	Clayton 270^o	-0.07	-
	9, 2; 12, 7	Frank	0.37	-
	7, 10; 12, 11	Clayton	0.10	-
	4, 9; 7, 12	Tawn type 2 180^o	1.11	0.09
	11, 4; 7, 12	Clayton 180^o	0.06	-
4	7, 6; 2, 8, 3	Frank	-0.26	-
	12, 8; 7, 2, 3	Independence	-	-
	2, 1; 9, 7, 12	Independence	-	-
	9, 3; 12, 7, 2	BB8	1.33	0.92
	11, 5; 7, 12, 4	Independence	-	-
	4, 2; 9, 12, 7	Independence	-	-
	4, 10; 7, 12, 11	Gaussian	0.08	-
	11, 9; 4, 7, 12	Independence	-	-
5	12, 6; 7, 2, 8, 3	Gaussian	0.18	-
	9, 8; 12, 7, 2, 3	Gaussian	0.11	-
	3, 1; 2, 9, 7, 12	Student t	-0.05	20.15
	4, 3; 9, 12, 7, 2	Gaussian	0.19	-
	10, 5; 11, 7, 12, 4	Independence	-	-
	11, 2; 4, 9, 12, 7	Independence	-	-
	9, 10; 4, 7, 12, 11	Gumbel 270^o	-1.05	-
6	9, 6; 12, 7, 2, 8, 3	Gaussian	-0.15	-
	4, 8; 9, 12, 7, 2, 3	Frank	0.45	-

	4, 1; 3, 2, 9, 7, 12	Independence	-	-
	11, 3; 4, 9, 12, 7, 2	Independence	-	-
	9, 5; 10, 11, 7, 12, 4	Independence	-	-
	10, 2; 11, 4, 9, 12, 7	Independence	-	-
7	4, 6; 9, 12, 7, 2, 8, 3	BB8	1.09	0.90
	11, 8; 4, 9, 12, 7, 2, 3	Independence	-	-
	11, 1; 4, 3, 2, 9, 7, 12	Independence	-	-
	10, 3; 11, 4, 9, 12, 7, 2	BB7 270^{o}	-1.05	-0.15
	2, 5; 9, 10, 11, 7, 12, 4	Gaussian	-0.06	-
8	11, 6; 4, 9, 12, 7, 2, 8, 3	Frank	-0.24	-
	10, 8; 11, 4, 9, 12, 7, 2, 3	BB8 90^{o}	-1.74	-0.67
	10, 1; 11, 4, 3, 2, 9, 7, 12	Frank	0.28	-
	5, 3; 10, 11, 4, 9, 12, 7, 2	BB1 180^{o}	0.24	1.04
9	10 ,6; 11, 4, 9, 12, 7, 2, 8, 3	BB1 180^{o}	0.12	1.03
	1, 8; 10, 11, 4, 9, 12, 7, 2, 3	Independence	-	-
	5, 1; 10, 11, 4, 3, 2, 9, 7, 12	Independence	-	-
10	1, 6; 10, 11, 4, 9, 12, 7, 2, 8, 3	Gumbel	1.02	-
	5, 8; 1, 10, 11, 4, 9, 12, 7, 2, 3	Independence	-	-
11	5, 6; 1, 10, 11, 4, 9, 12, 7, 2, 8, 3	Gaussian	0.04	-

Notes: The number coding is identical to the one used in Figures D.1-D.11 in Appendix D. The semicolon separates the variables (left) for which the conditional association is modelled from the conditioning variables (right).

CHAPTER 2: SELF-REPORTED LIFE SATISFACTION THROUGH THE LENS OF TREE-BASED LONGITUDINAL ANALYSIS

Abstract: The main aim of this study is to examine the determinants of well-being through employing a machine learning technique. Well-being is measured by self-reported life satisfaction from the UK household longitudinal survey. The technique used is the RE-EM tree by Sela and Simonoff (2012). The proposition made in this study intends to complement the standard linear techniques often used in the literature to provide a more comprehensive perspective. Machine learning methods require no *a priori* structure or variable selection. In the complex and inclusive context of life satisfaction this can prove useful as several non-linearities and interactions between covariates, that would otherwise seem unlikely to be pre-specified, reveal themselves to the researcher. The well-being structure suggested by the RE-EM tree is compared to a linear model. The explanatory power of the two is comparable suggesting that the non-parametric estimation can offer useful insights and complement the traditional parametric approach. The estimated RE-EM tree structure replicates many of the results in literature with regard to major determinants of well-being after implementing a predictive margins postestimation analysis. At the same time, it allows the classification of individuals into different well-being groups according to a combination of their individual characteristics. This grouping can be helpful in deriving targeted policies and interventions.

1. INTRODUCTION

Life satisfaction is one of a set of variables associated with facets of well-being which are inherently unobserved. Other intuitive components of this set include happiness, anxiety, and mental health. Given the plethora of interesting questions that can be addressed using these concepts, there is a need to ascertain the most appropriate way to include these concepts in quantitative analysis. One possible way is the use of subjective, self-reported measures. These measures can act as proxies for the true unobserved values, thus providing a 'tangible' quantity that can be used in analytical settings, such as regression function estimation.

A strand of the literature deals with the accuracy, and thus usefulness, of such measures in capturing what they aim to record (see for example, Bertrand and Mullainathan, 2001, Bond and Lang, 2019, and Oswald and Wu, 2010). Another strand of the literature assumes that the measures are adequate proxies and simply investigate the structural form of their data-generating processes (see for example, Gerdtham and Johannesson, 2001, Clark and Oswald, 2002, and Boyce *et al.*, 2010). In the context of life satisfaction it is assumed self-reported measures used to capture life satisfaction do indeed encapsulate the actual, unobserved value of the variable for each individual in a satisfactory manner. Based on this assumption, focus is given to understanding the determinants of life satisfaction.

A significant fraction of the studies to date that use life satisfaction, and other well-being notions, employ parametric methods which require pre-specifying, and thus assuming, the functional form of life satisfaction. For example, many assume that life satisfaction depends linearly on a set of pre-selected explanatory variables. This implies that an additive structure is assumed prior to estimating the implied model, in which every variable acting as a determinant of life satisfaction is known. Ferrer-i-Carbonell (2013), and Clark (2018) offer comprehensive reviews of the findings associated with such parametric approaches. Several important results have emerged from the use of such approaches over the past four decades. Examples include the U-shaped association of well-being with age, the significant impact of social comparison on well-being, as well as the adaptation of well-being across time to events that initially may have had a significant impact on individual welfare.

The main aim of the current study is to examine the determinants of self-reported life satisfaction through employing a tree-based methodology. The motivation is that tree-based methodologies allow for both non-parametric estimation of the association between the

dependent variable and covariates, and variable selection[1]. The proposition made in this study does not intend to offer a substitute to how well-being should be studied, but rather a complement attempting to provide a more comprehensive perspective.

Tree-based methods have a data mining nature, and as such require no structure, and no *a priori* variable selection. In the complex and inclusive context of life satisfaction this can prove useful as several non-linearities and interactions between covariates, that would otherwise seem unlikely to be pre-specified, might reveal themselves to the researcher.

Tree-based methodologies belong to the general class of statistical learning techniques (Hastie *et al*., 2009), with the most famous example being the classification and regression trees by Breiman *et al.* (1984). One of the characteristic features of tree-based procedures is the partitioning of the sample into non-overlapping regions and, subsequently, the use of the dependent variable's (i.e. self-reported life satisfaction in this case) empirical distribution within each region to make comparisons across regions. The RE-EM tree is a version of tree-based procedures which respects the longitudinal nature of the data if present[2].

Despite their popularity, machine-learning techniques, which nest tree-based procedures, have only made a modest entry in the applied economics literature. Varian (2014), and Mullainathan and Spiess (2017) give a brief description of how such techniques can be beneficially used by economists. As far as the well-being literature is concerned, very few studies have attempted to make use of the non-parametric tree-based approaches. Exceptions include, Galletta (2016) who makes use of a classification tree to explore the determinants of a modified happiness dummy variable; and Morrone *et al*. (2019) who propose a classification tree-based technique that encompasses the ordinal nature of a life satisfaction variable. Both of these studies emphasize the benefits of using a tree-based methodological approach when studying well-being by highlighting the new insights that can arise.

It should be noted that the author of the thesis was made aware of a working paper on machine learning by Oparina *et al*. (2022) during the viva examination. The aforementioned study assesses the potential of machine learning techniques in understanding well-being and finds that such approaches tend to have higher predictive power in relation to traditional models. Just

[1] Variable selection in the sense that the explanatory variables used for estimation are automatically selected from a pre-specified set of variables. The main advantage is that the pre-specified set can have as many variables as desired, even if that is more than the number of observations in the data set, without necessarily running into the issue of overfitting like in the case of e.g. a linear regression.

[2] In particular, the estimated model under the RE-EM tree procedure incorporates individual-level effects which are common to repeated observations from the same subject.

like in the current thesis, the authors also find that results from machine learning approaches validate the ones from conventional techniques.

This study aims to contribute to the well-being literature by both extending tree-based analysis to incorporate the longitudinal structure of data used in the RE-EM tree, and providing a link with existing parametric approaches by presenting the estimated tree in a linear model. The studies mentioned above use tree-based approaches for cross-sectional analysis. As with any type of methodology, longitudinal analysis can offer new insights to the subject under examination, and for life satisfaction it is probably a more realistic approach to modelling. This is done by using seven waves from Understanding Society[3], UK's household longitudinal survey, spanning from 2010 to 2018, and the RE-EM tree approach by Sela and Simonoff (2012). Another aspect of this study that is usually neglected in the literature is the inclusion of personality traits in the analysis of life satisfaction determination. Measures provided in Understanding Society will be used to approximate individuals' personality characteristics, an inclusion proposed by Ferrer-i-Carbonell and Frijters (2004) given their importance for "general satisfaction".

The explanatory power of the estimated well-being tree is similar to that of a linear model which uses the same input as the RE-EM tree. In addition, the estimated tree structure replicates many of the results in the well-being literature when implementing a predictive margins postestimation analysis. The non-parametric estimation offers additional insights in that it allows classifying individuals into different well-being groups according to a combination of their individual characteristics. This is possible due to the various interactions generated during the estimation of the tree, a feature inherent to tree estimation. This grouping can be beneficial in designing targeted policies.

The following section will provide a literature review associated with both the importance of well-being measures, and results regarding the determinants of life satisfaction and other well-being notions. Subsequent sections will present the Understanding Society data set used in the analysis, the RE-EM tree methodological approach in more detail, results based on the estimation of the RE-EM tree, a comparison with the within estimator, a parametric linear approach, and postestimation analysis outcomes in the form of predictive margins.

[3] Available on https://www.understandingsociety.ac.uk/.

2. LITERATURE REVIEW

2.1 Subjective, self-reported measures

2.1.1 Significant overlap of literature with other chapters

Despite being extensively used, subjective, self-reported measures are not unanimously considered to be bulletproof proxies for life satisfaction and other unobserved variables. Bertrand and Mullainathan (2001) question subjective measures' validity based on the argument that things like cognitive factors and social desirability can influence individuals' responses. Some examples of cognitive effects mentioned by the authors include the influence of the order and the wording of the questions in a survey on the responses given by individuals. Experimental evidence has shown that people's responses can vary as a result of changing the order or wording of subjective questions; and that the social desirability factor can influence an individual response in a way other than what they believe to be true to avoid seeming 'bad' in front of the interviewer. One example is the case of asking the questions "How happy are you with life in general?", and "How often do you normally go out on a date?". When the question regarding dating was first, there was high correlation between the responses of the two. When it was the other way around, the two were essentially uncorrelated. Using a measurement-error framework, the authors suggest that there might be some value in using such measures as predictors, but their use as dependent variables is inappropriate. These issues raise concerns about the validity of subjective measures as they demonstrate how responses to the relevant questions can be influenced by the survey procedure, which should be irrelevant.

Another concern raised is that, without imposing strong auxiliary assumptions, it is impossible to compare groups of individuals based on the estimated mean values of their true underlying happiness distributions by using survey data. Bond and Lang (2019) operate in a context in which happiness is considered equivalent to the notion of utility, and therefore there can be infinite candidates for the true underlying happiness distribution that can preserve the choices observed to be made by individuals on the ordinal scale provided to them[4]. As such, the only way to make sure that the mean ranking of groups remains the same for all possibilities is to establish that the happiness distribution of one group first order stochastically dominates (FOSD) the other. The authors propose that it is highly unlikely for the conditions of non-parametric identification of stochastic dominance to be met in practice (e.g. groups cannot be ranked in terms of FOSD if both have observed responses in the highest and lowest categories

[4] This is similar to the idea that any monotonic transformation of a utility function represents the same preference ordering. In this case the ordering is concerned with states of happiness.

of the survey's ordinal scale). When it comes to parametric identification, the authors suggest that it is almost impossible to establish stochastic dominance in the case that the underlying distributions of the groups come from the same unbounded location-scale family. The reason behind this is that arbitrary monotonic transformations of the scale can reverse the mean ranking of groups. The only chance of identifying stochastic dominance is in the unlikely case of the equality of variances of the happiness distributions between the groups under consideration. Despite the above, they still recognize the possibility for a commonly accepted consensus on the structure of the true happiness distribution that would allow for meaningful analysis[5].

From the perspective of the psychology literature, the mood of the individual at the time of responding to the relevant survey questions can influence their evaluation of happiness and satisfaction with their lives. Based on two experiments, Schwarz and Clore (1983) demonstrate how the information about the true underlying value captured by the reported measure can be confounded by the momentary mood as shaped either by incidents associated to the individual, or by external factors such as the amount of sunshine.

A number of authors argue in favour of subjective, self-reported measures. Alesina *et al*. (2004) offer a comprehensive review of why they can be considered reliable. They base their argument on findings such as the association of reported happiness with more objective measures such as blood pressure, heart rate, and prefrontal brain activity. The authors also mention the possibility that the influence of social desirability on individuals' responses is exaggerated, based on evidence by psychologists[6]. Oswald and Wu (2010) also argue in favour of the validity of subjective measures by showing how they are strongly correlated across geographical areas with objective measures constructed from non-subjective data in a compensating-differentials approach. By using data from the U.S. Behavioral Risk Factor Surveillance System[7], the authors construct regression-adjusted[8] life satisfaction estimates for 50 U.S. states. These estimates, representing the subjective version of measuring life satisfaction, are found to have a strong association with an objective quality-of-life ranking for the states, constructed by Gabriel *et al*. (2003) based on indicators measuring aspects such as

[5] Their conclusions can be extended to incorporate any other unobserved variable for which the distribution is approximated by subjective responses to survey questions recorded on some form of ordinal scale.
[6] Studies by Rorer (1965), and Konow and Earley (1999) are mentioned in support of the argument.
[7] Available on https://www.cdc.gov/BRFSS/.
[8] Adjusted by controlling, among other things, for income, age, gender, ethnicity, education, marital, and employment status.

sunshine, temperature, violent crime, air quality, student-teacher ratio, taxes, and many other features of life.

Oswald (2008) attempts to provide a conceptual framework within which the association between subjective measures and their true, unobserved counterparts can be expressed. The author makes use of the notion of a reporting function that maps the values of objective, internal feelings for an individual to the reported values given as responses to survey questions. In an attempt to estimate the reporting function, Oswald (2008) uses a quasi-experimental design in which participants are asked to report subjective evaluations for their height relative to individuals of their own gender[9], a variable which can be objectively measured as well. Based on their responses, a regression estimation suggests an increasing reporting function which exhibits slight concavity. Partitioning the sample based on gender and re-estimating the regression equation for each partition suggests linearity of the reporting function for both genders. Under the assumption of the existence of a reporting function, the first derivative and the second derivative of the function[10] are of vital importance in comprehending the link between reported and true, unobserved values. In the case that the reporting function is common to all and exhibits linearity, the most restrictive assumption in the use of subjective measures as stated by Ferrer-i-Carbonell and Frijters (2004), the one of cardinal interpersonal comparability, is satisfied. The evidence for heterogeneity of the reporting function across groups of individuals might not be in favour of using subjective measures as proxies for the true values because interpersonal comparison without any prior information about the heterogeneity in reporting could be misleading. However, it should be noted that the quadratic term in the estimation based on the whole sample is marginally significant, contributing little explanatory power[11], as opposed to linearity.

Assuming that subjective measures are adequate proxies for quantitative analysis, another question addressed in the literature is to ask what type of subjective well-being measure should be used and how it should be treated. Clark (2015) considers three of the main types, including life satisfaction, affect, and eudaimonia[12], and finds that they are significantly correlated with each other, as well as being associated with certain explanatory variables (e.g. education,

[9] Individuals are asked to report how tall they feel relative to individuals of their own gender by recording their answer on a continuous line ranging from 0 being the case of *very short* to 10 being the case of *very tall*.
[10] Implicit assumption of the reporting function being continuously differentiable.
[11] On the basis of R^2.
[12] Affect has to do with an instantaneous judgment of how an individual is feeling. Eudaimonia is a concept dealing with an individual achieving potential in various aspects of life.

marital status, log income, etc.) in approximately the same way. However, the author notes that there are examples in literature where this finding might not hold, especially if experienced and evaluative measures for well-being are compared. Experienced well-being refers to emotions as experienced by individuals on a day-to-day basis, whereas evaluative well-being refers to the well-being individuals derive from overall evaluations of their lives. Furthermore, he states that the choice between ordinality and cardinality as an assumption for the nature of the measures is relatively inconsequential for the conclusions drawn from the analysis. The author's argument is based on the strong correlation exhibited between the estimated coefficients of estimators which assume ordinality of the dependent variable (ordered probit), and the estimated coefficients of estimators which assume cardinality (ordinary least squares). Similar findings are presented by Ferrer-i-Carbonell and Frijters (2004), and Pfaff (2013). Another view on the interrelation of different concepts is given by Tomer (2011) who proposes that happiness could be broken down such that $Happiness = S + U_C + E$, where S represents an individual's set point, U_C reflects the contribution by hedonic features of life, and E stands for eudaimonia.

Despite the substantial amount of work on understanding how the measures interplay with each other, many still aim to discover the underlying determinants of well-being by neglecting the type of measure used. For example, Easterlin (2005) assumes that terms like happiness, life satisfaction, and well-being are interchangeable in his attempt to provide a general theory for well-being.

Overall, despite the concerns about subjective, self-reported measures being suitable for quantitative analysis, there is encouraging evidence that they can be viewed as adequate proxies. Furthermore, there is still the need for a universal framework of well-being that embraces the different concepts, such as happiness, and life satisfaction. Despite the 'loose ends' surrounding the literature concerned with subjective measures, the appeal for ways to incorporate welfare measures that go beyond the standard economic indicators (e.g. GDP) should be uncontroversial. Diener and Seligman (2004) suggest that economic indicators have shortcomings when it comes to representing the wants and needs of current societies, and that they were more relevant during the initial stages of economic development. They propose the use of more inclusive well-being measures that would be policy relevant. In a similar spirit, the Stiglitz Commission (2009) presents shortcomings of measures based on economic performance and proposes the possibility of going beyond GDP towards a multi-dimensional

view of well-being. Sooner or later, a common consensus on how to define and measure well-being will be established.

2.2 Well-being determinants

2.2.1 Significant overlap of literature with other chapters

Even with a considerable part of the literature dealing with the subjective measures' usefulness per se, many studies take this as given and go on to examine the determinants of well-being by laying out welfare equations where the dependent variable is self-reported. Various concepts are used across the different studies, including happiness and life satisfaction. One of the first studies by Gerdtham and Johannesson (2001), using Swedish microdata, finds that happiness is associated with higher levels of income, health, and education for an individual. On the other hand, being unemployed, male, and single is negatively related to happiness. In addition, they are among the first to suggest a U-shape association between a well-being notion and age, a finding which has received significant attention over the years[13]. Clark and Oswald (2002) confirm some of the aforementioned findings through their analysis which also aims at assigning a monetary compensating value to such life events based on relative coefficient values estimated for their happiness regression equations. The value of the estimated coefficient for the life event of interest is used along with the estimated coefficient for the income variable. The relative value between the two provides an indication of the monetary value that could have the same influence on happiness as the life event.

Studies have also considered the effect of social comparisons. The study by Clark and Oswald (2002) is also part of the literature which provides evidence for the significant effect of social comparisons on well-being, in the sense that relative and not only absolute quantities (e.g. absolute income) can matter for individuals' happiness. Using a reference income specification, i.e. using the idea that individuals have a particular set point to which they compare their income, studies such as Ferrer-i-Carbonell (2005), and Becchetti *et al.* (2013), propose a negative association of reference income with happiness and life satisfaction respectively. Such a finding suggests that individuals experience adverse feelings when they fall short of what their comparison benchmark is. It may also be the case that reference income exerts a positive

[13] Blanchflower and Oswald (2008), Glenn (2009), and Frijters and Beatton (2012) are only some of the studies arguing whether well-being exhibits convexity across the lifespan of individuals. One of the main challenges in providing a definitive answer comes in the form of disentangling the impacts of age, period, and cohort on well-being. Due to perfect collinearity, only two out of the three can be included in a linear specification.

influence on life satisfaction because of the possibility that it contains an information component with regard to the future economic progress of individuals (Senik, 2004).

An alternative proposition for the determination of well-being is the idea that social comparison takes the form of an ordinal ranking of individuals with respect to income. For example, a study by Boyce *et al*. (2010) demonstrates how this type of specification is dominant[14] as opposed to specifications which incorporate a cardinal measure of own income and reference income for life satisfaction. Similar findings have been exhibited when using other notions of well-being as the dependent variable in the analysis. The dominance of the specification using ordinal ranking has been shown when examining the determinants of mental distress (Wood *et al*., 2012), health (Daly *et al*., 2015), and even depressive symptoms (Osafo Hounkpatin *et al*., 2015).

The use of personality traits in well-being analysis is supported by Ferrer-i-Carbonell and Frijters (2004) who propose the use of time-invariant personality characteristics as candidate regressors that can account for the fixed unobserved components which usually exist in a model aimed at capturing the data-generating process of data sets with panel structure. Borghans *et al*. (2008) offer a more general review of how the economics and psychology of personality traits could be usefully integrated. For example, they suggest that psychology findings concerning personality should be used to inform economic models, as well as econometric methods being used to study the formation and evolution of personality traits. Introducing personality traits, Proto and Rustichini (2015) find that the level of neuroticism can significantly influence the effect of income on life satisfaction. Based on this finding, there is evidence of interaction between personality traits and other variables in the determination of life satisfaction. Evidence for interactions of this nature demonstrate the possible complexity in well-being determination.

Ferrer-i-Carbonell (2013) also refers to the evident process of adaptation with respect to the influence of life events on well-being. Adaptation can refer to the re-adjustment of aspirations when a certain target level is achieved. For example, the case of income rising does not necessarily imply a long-term rise in well-being as now an updated, higher reference point may enter the well-being determination. Ferrer-i-Carbonell and Van Praag (2008) use a long German household panel data set to provide evidence that the adaptation of life satisfaction to

[14] Dominance in this case is established in terms of a significance comparison of the estimated coefficients for the relevant variables when they are simultaneously included in the specification.

income changes is partial, meaning that income changes may have a long-term influence. They also suggest the possibility for asymmetry in the magnitude of the influence of income changes on life satisfaction depending on the direction of the change. In contrast, Di Tella *et al*. (2010) using a sample from the same German panel survey find evidence in favour of happiness adapting totally to income changes over the long term. When it comes to health, Oswald and Powdthavee (2008) propose that even in the case of severe disability individuals demonstrate a partial adaptation of mental well-being. Using the British Household Panel Survey, they estimate a degree of adaptation varying from 30% up to 50% depending on how severe the type of disability is[15].

2.3 Tree-based methodology

Before looking into the literature which uses tree-based approaches to analyse well-being, it is useful to provide a brief introduction into how the interpretation of such methods is approached. A more detailed explanation is provided in section 4. Trees are constructed based on a process of sequential binary splitting of the sample at hand. The criterion for splitting a (sub)sample is determined by the segregation of the values in the domain of one of the explanatory variables included in the estimation such that the two resulting subsamples have (almost always) different average predictions for the value of the dependent variable[16]. As such, the splitting of the (sub)sample depending on the values of some explanatory variable along with the resulting average prediction of the dependent variable can indicate some form of association between the two. The sample may be partitioned several times before the final tree structure is determined. In addition, some explanatory variables may be used repeatedly in the construction of the tree, while others may not be used at all, depending on what is optimal in terms of the relevant objective function.

In one of the few studies analysing well-being through tree-based approaches, Galletta (2016) finds a positive association of happiness with income and financial assets. This is a finding that mostly agrees with the literature in the sense that the coefficient indicating the association between income and self-reported satisfaction is almost always found to be positive and statistically different from zero (Ferrer-i-Carbonell, 2013). However, the study by Galletta (2016) introduces no measures associated with the social comparison of income and thus

[15] An inclusive report of the numerous findings in well-being determinants research is given by Ferrer-i-Carbonell (2013). Apart from the variables pointed out above, she also presents evidence on the influence of individual characteristics such as religion, political beliefs, children, and obesity, as well as aggregate regional characteristics such as inflation, unemployment rate, and GDP.

[16] In a manner that optimises some objective function (e.g. residual sum of squares).

cannot be directly compared with more comprehensive studies, such as the one by Boyce *et al*. (2010). In addition, being married appears to have a consistent positive association with happiness throughout the tree presented by Galletta (2016). It is, however, the case that the estimated tree indicates heterogeneity between individuals in terms of the association of the aforementioned explanatory variables with happiness. This heterogeneity is determined by the interaction of explanatory variables included in the estimation. The substantial and informative level of non-linearity indicated by the tree without any model imposed *a priori* is one of the main reasons to use this approach to complement well-being empirical analysis. Galletta's (2016) study is based on Italian cross-sectional data and uses a binary dependent variable by modifying the original variable recorded on a 10-point scale.

Morrone *et al*. (2019), using their novel methodological approach, demonstrate a negative impact on life satisfaction for individuals who are economically disadvantaged[17], as well as substantially low life satisfaction for those who are unemployed. Both of these findings seem to agree with the previous literature outlined above. The authors derive a new approach to generate trees which takes into account the ordinality of the life satisfaction measure they use, instead of assuming cardinality. The novelty here comes in the form of the weaker ordinality assumption as opposed to the more demanding one of cardinality when it comes to the subjective dependent variable, while still using a tree-based approach. Another aspect of life which appears as important in their tree structure is the quality of relationships that individuals have in their lives, both with respect to their families, and with respect to their friends[18]. More 'satisfying' relationships seem to be associated with higher levels of life satisfaction. The interactions between the covariates in the estimated tree act as an indication of the possible structure of non-linearity that has to be kept in mind when investigating life satisfaction. In this particular case, it comes mainly in the form of the interaction between economic conditions and the quality of relationships, influencing the predicted level of life satisfaction.

2.4 Following steps

Based on the approaches and results of the two aforementioned studies, it can be argued that there is the basis and room for further exploration of well-being concepts using such non-parametric approaches. The present study proposes to add value to the existing tree-based studies by using an extension of the standard regression tree by Breiman *et al.* (1984) which

[17] Due to the lack of an income variable, the authors use objective and subjective proxies to indicate economic conditions for individuals.

[18] The authors use the 2014 wave of ISTAT's Multipurpose Survey on Everyday Life Aspects which asks questions about social relationships, permitting them to perform the relevant analysis.

accounts for the longitudinal structure of the data. Breiman *et al.* (1984), who provide both practical and theoretical sides to the use of tree methods, also suggest the use of trees as complementary non-parametric tools rather than substitutes to other approaches. The study will also take a concept introduced by Ferrer-i-Carbonell (2013) by investigating the inclusion of personality characteristics in well-being analysis through non-parametric estimation. The intention is to compare the existing parametric approaches with this alternative non-parametric method to provide new insights and aid understanding when using these types of measures. Along with parametric methods, the two main features of tree-based methods that can contribute to a better understanding of the association of other variables to life satisfaction are the non-parametric estimation, and variable selection. Like in the studies above, the non-parametric estimation can reveal interactions and non-linearities in how explanatory variables are associated with well-being that would otherwise seem unlikely to be pre-specified. In addition, the variable selection feature allows the incorporation of a large number of variables in the analysis as the estimation procedure only chooses the most relevant ones.

3. DATA

3.1 Understanding Society

The sample used to estimate the RE-EM tree in this study comes from seven waves of Understanding Society, the UK's household longitudinal survey, spanning from 2010 to 2018[19]. This sample consists of 264,518 observations from 64,260 individuals. Summary statistics of the main variables used are provided in Appendix A. Understanding Society is a multi-purpose nationally representative survey of British households, including extensive socio-economic and psychological modules. It incorporates samples from England, Wales, Scotland, and Northern Ireland. The panel nature of the data set allows for the use of methodologies, including the RE-EM tree, that can take advantage of observations nested within the same individual.

3.2 Life satisfaction index

The dependent variable in the subsequent analysis is life satisfaction. As mentioned in the introduction, this variable cannot be observed, and thus a subjective, self-reported measure is used as a proxy instead. The variable used comes from individuals being asked to assess their life overall as a response to the question:

"Here are some questions about how you feel about your life. Please choose the number which you feel best describes how dissatisfied or satisfied you are with the following aspects of your current situation: Your life overall."[20]

The responses to this question are recorded on a 7-point Likert scale ranging from 1 being *"Completely dissatisfied"* to 7 being *"Completely satisfied"*.

Whilst acknowledging the limitations of such self-reported measures, as discussed in the literature review, for this analysis we assume the measure to be an adequate proxy. This study's aim is to demonstrate the added value of using a non-parametric approach. As such, it abstracts from considerations of measure inadequacy. Given that the assumptions associated to the self-reported measure in this study are similar to a significant portion of the literature, any new insights generated can be attributed to the methodological component of the paper.

[19] The first wave of Understanding Society is not incorporated in the analysis as it does not record the variable associated to health, which turns out to be very important in terms of its explanatory power for life satisfaction.
[20] Questionnaires available on https://www.understandingsociety.ac.uk/documentation/mainstage/questionnaires.

Based on the point made by Ferrer-i-Carbonell and Frijters (2004), Pfaff (2013), and Clark (2015), the life satisfaction variable is assumed to have a cardinal nature regardless of the fact that it is recorded on an ordinal scale. As the authors above suggest, the distinction between ordinality and cardinality can be relatively inconsequential for the conclusions drawn from the analysis.

3.3 Life satisfaction determinants

Drawing on the existing literature, the variables incorporated in the analysis as independent variables can be divided into two categories. Those that measure standard socio-economic characteristics; and those that measure personality traits.

Socio-economic characteristics, recorded at the individual level, include age, and the natural logarithm of equivalised household income[21], as well as sets of dummies for economic activity, country of residence, gender, marital status, highest educational qualification, general health[22], number of own children in household, and ethnicity. In addition, year dummies are included based on the calendar year during which the interview was carried out.

The other set of covariates used are the variables aimed at capturing the personality characteristics of the individuals interviewed. The variables available in Understanding Society represent the Big Five personality traits. Studies such as that of Goldberg (1990), and McCrae and John (1992) are supportive of the general applicability of this method of capturing the overall structure of an individual's personality. Goldberg (1990) bases the support on the argument that the analysis of any rich sample of English adjectives describing an individual's traits (either own or peer's) will elicit some Big Five variant. One of the main appeals, suggested by McCrae and John (1992), in support of using the Big Five's dimensions as predictors is the construct's comprehensiveness in systematically representing the most important aspects of personality. The authors promote Big Five's replication across cultures, and empirical validation across methodological approaches. The five dimensions describing an individual's personality include extraversion (e.g. being outgoing and talkative), agreeableness (e.g. being trusting and kind), conscientiousness (e.g. being responsible and thorough), neuroticism (e.g. being anxious and worrying), and openness (e.g. being creative and curious).

[21] Pfaff (2013), in a study aiming to explore the features involved in the analysis of life satisfaction when using survey data, promotes the use of equivalised household income as it accounts for household size and composition. Therefore, the square root scale is used (OECD, 2011). In addition, the income variable is adjusted for inflation so that it represents real income. UK inflation data is available by the Office for National Statistics on https://www.ons.gov.uk/economy/inflationandpriceindices.

[22] Health assessment is based on a subjective, self-reported measure.

Personality characteristics are captured only once in wave 3 of the Understanding Society survey out of the seven waves used in this study. As such, in order to preserve the longitudinal structure of the data set, the assumption of stability of the personality characteristics is made. Such an assumption may not be implausible as demonstrated by Roberts and DelVecchio (2000) who find that personality traits are quite consistent over the life span of individuals. Theirs is a meta-analytic study using 152 longitudinal studies and is based on the concept of rank-order consistency. The variables representing each of the five dimensions are recorded on a 7-point Likert scale. For the subsequent analysis, these variables are assumed to be ordinal in nature, and are thus included as a set of dummies in the specification. The main reason for including them as sets of dummies is the possibility of capturing any non-linearity in the association between personality traits and life satisfaction. For example, in a linear regression, a set of dummies would be flexible enough to capture a quadratic association as each dummy is allowed to shift the intercept accordingly. However, including a continuous variable instead of dummies would force the estimator to a linear association. Preliminary analysis of the association between life satisfaction and personality traits is presented in the next subsection.

3.4 Personality traits

The construction of the five personality dimensions is based on the reported values of fifteen survey questions (three per dimension), also recorded on a 7-point Likert scale. The value for each dimension is given by the rounded average of the responses of individuals to the three survey questions corresponding to each dimension. Each survey question corresponds to an attribute for which the individual responding is required to make a self-evaluation and then report it on the 7-point scale. *Table 1* shows the personality dimensions along with their associated attributes.

Table 1: Personality dimensions and associated attributes.

Personality dimension	Attribute
Extraversion	Talkative
	Sociable
	Reserved
Agreeableness	Rude
	Forgiving nature
	Kind
Conscientiousness	Does a thorough job
	Lazy
	Efficient
Neuroticism	Worries a lot
	Nervous
	Relaxed
Openness	Original
	Artistic
	Active imagination

A preliminary analysis of the association of personality traits with each other, as well as with life satisfaction, is presented here in the form of a partial correlation network. For a set of variables, partial correlation denotes the linear association between any two variables after conditioning on the rest of the variables in the set (Cox and Wermuth, 1993). For a generic vector of variables $\boldsymbol{x}$, let $\boldsymbol{\Sigma}$ denote the positive definite covariance matrix of the variables. $\boldsymbol{\Sigma^{-1}}$ represents the concentration matrix, the inverse of the covariance matrix. σ_{ij} denotes the element in row i and column j of the concentration matrix. Based on the specified notation, the partial correlation ρ_{ij} between any two variables i and j is given by:

$$\rho_{ij} = -\sigma_{ij}\left(\sigma_{ii}\sigma_{jj}\right)^{-\frac{1}{2}}.$$

For the set of variables generated by the union of the fifteen personality attributes and life satisfaction, the sample counterpart of partial correlation is calculated for every pair of variables in the set based on the aforementioned formula. As such, a symmetric partial correlation matrix can be specified, also known as the weights matrix (Costantini *et al.*, 2015).

The weights matrix acts as the basis for generating a network representation of the association between the variables of interest. In a network, the individual entities are represented by nodes. In this case, nodes represent individual variables. The bilateral associations between nodes in a network are given by links between nodes. Associations based on the weights matrix generate undirected, symmetrical links as they denote partial correlation values. Therefore, the existence

of a link denotes the existence of some form of conditional linear dependence between the linked variables, as opposed to being conditionally linearly independent in this case. The partial correlation network formed based on the weights matrix is given by *Figure 1*.

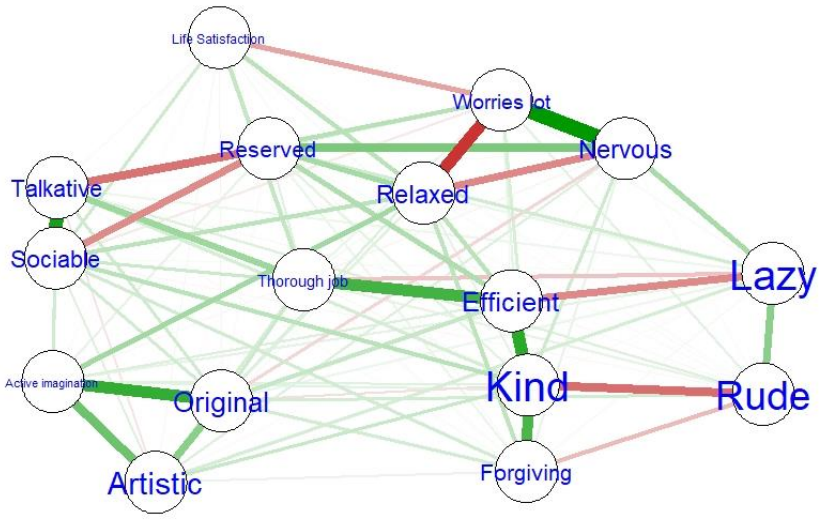

Figure 1: Partial correlation network.

Notes: Green links account for positive association and red for negative. The brightness and thickness of the links account for the strength of association. Strength is measured in relative terms with the strongest association providing the reference point.

As mentioned, the nodes in *Figure 1* represent the sixteen variables incorporated in the preliminary analysis. Green links account for positive association, whereas red for negative. The brightness and thickness of the links account for the strength of conditional linear association. The figure is such that the strength is measured in relative terms with the strongest association providing the reference point. The exact partial correlation values in the weights matrix are provided in Appendix B.

With the visual aid of the partial correlation network, it is clear that the strongest associations exist between the attributes linked to the same personality dimension. This is a reassurance of the good quality of subjective responses given by individuals. What is worth noting is the fact that life satisfaction does not exhibit relatively strong links with any of the attributes associated with personality, at least in the context of conditional linear dependence. This can act as an indication that there is significant variation in life satisfaction that may be explained by factors other than personality.

Apart from the visual aid that can be offered by the network representation of partial correlation, there are also features inherent to the notion of networks that can be used to

understand the role of individual entities within the specified network. One such feature is the individual node property known as strength (Costantini *et al*., 2015). Strength is a node property that attempts to measure how central a particular node is for the network in consideration. In particular, node strength is defined as the sum of the absolute values associated with the node's direct links to other nodes. *Figure 2* is a plot of the sixteen strength values associated to each variable.

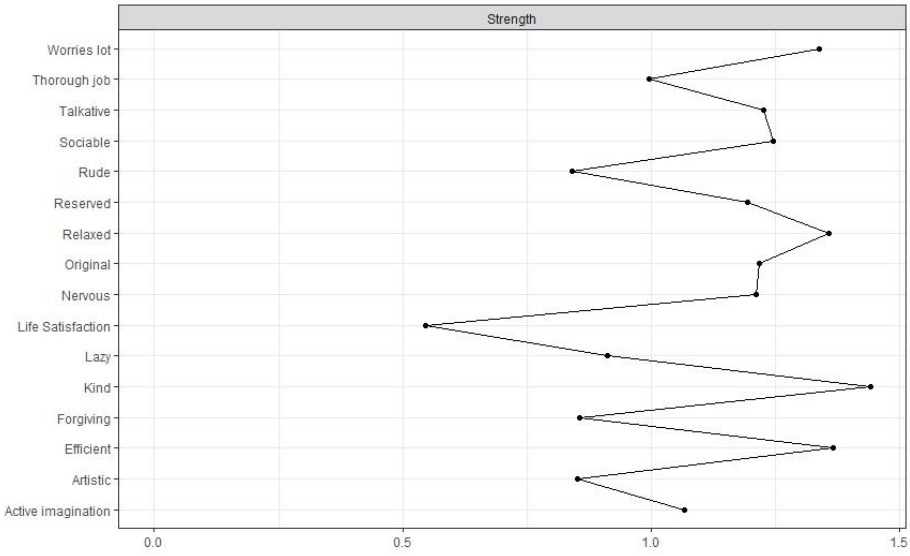

Figure 2: Node strength.

Life satisfaction stands out as the least important variable based on node centrality. This confirms in a more formal manner the observation made previously based on the partial correlation network in that life satisfaction exhibits relatively weak direct links with the personality attributes. This provides motivation for further exploration of variables, other than the ones accounting for personality, which can be used to explain life satisfaction determination.

4. METHODOLOGY

4.1 Regression trees

Before presenting the RE-EM tree methodology, an introduction to the notion of tree-based estimation through regression trees (Breiman *et al*., 1984) will be outlined based on James *et al*. (2013). Regression tree estimation is part of the RE-EM tree estimation procedure.

A first step to appreciating the reasoning behind the use of regression trees comes from understanding the limitations of more well-established techniques, such as the case of a linear regression. Using the classical linear regression model (CLRM) to represent any data-generating process assumes that each variable from a set of explanatory variables has a separate, additive effect on the dependent variable. There is, however, the possibility that this data-generating process exhibits 'severe' non-linearities. CLRM can accommodate non-linearities as long as they come in the form of pre-specified interactions between explanatory variables. Interactions can occur between two or more variables. However, each additional pre-specified interaction included in the model is associated with a higher number of parameters that need to be estimated, usually making it impractical to consider all intuitively reasonable interactions, let alone checking for every possible one. A regression tree offers a more 'natural' way of examining the interactions between explanatory variables by allowing them to emerge if they are important enough in the determination of the dependent variable.

A regression tree method splits the space generated by a vector of explanatory variables into non-overlapping regions. Therefore, depending on its combination of explanatory variable values, each observation is a member of a single region. Within each region there is a subsample of observations which can be used to construct an empirical distribution for the dependent variable of interest. The mean of the empirical distribution can act as a prediction for the value of the dependent variable for each observation within the region. The generation of regions is based on a process known as recursive binary splitting. Starting with the whole sample, sequential binary partitioning is performed based on a splitting criterion[23], which can be the choice of a threshold value based on a particular (continuous) explanatory variable such that the largest reduction in the residual sum of squares (*RSS*) is generated. The recursive partitioning is performed until a specified stopping criterion is achieved[23]. The variable selection feature is fundamental for the process as only the relatively most important predictors, in terms of the sequential *RSS* reduction contribution, may be used. For example, based on the

[23] Both the splitting and the stopping criteria can be chosen by the researcher. However, the latter is usually chosen adaptively based on a process known as cross-validation which is described further down.

aforementioned splitting criterion, a variable may be used more than once if it is optimal in terms of the *RSS* reduction; and since there is a stopping criterion it may be the case that not all variables are used in the construction of the estimated regression tree before meeting the criterion.

More formally, drawing on James *et al*. (2013), the recursive binary partitioning used to generate a regression tree can be summarised as:

$$\min_{j,s}\left[\min_{c_1}\sum_{\boldsymbol{x}_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{\boldsymbol{x}_i\in R_2(j,s)}(y_i-c_2)^2\right].$$

We are looking to find the splitting variable j and splitting point s to minimize the *RSS* and thus generate regions $R_1(j,s)$ and $R_2(j,s)$. $\boldsymbol{x}_i$ represents a vector of explanatory variables, and y_i the dependent variable of interest for observation $i \in \{1,\dots,N\}$. Splitting variable x_{ij} and splitting point s are such that:

$$R_1(j,s)=\{\boldsymbol{x}_i|x_{ij}\le s\} \text{ and } R_2(j,s)=\{\boldsymbol{x}_i|x_{ij}>s\}.$$

For a categorical covariate, a segregation of the values in the support of the variable is performed instead. As such, any combination of the values that the categorical variable can take, apart from the combination consisting of all the values, may end up in the same partition[24].

The inner minimization in the recursive partitioning specification above consists of choosing the values for constants c_1, c_2 which generate the smallest *RSS*. Therefore, given that $N_m = \sum_{\boldsymbol{x}_i\in R_m(j,s)} 1$ for a generic region $R_m(j,s)$, the mean of the empirical distribution in each region solves the inner minimization:

$$\hat{c}_1=\frac{1}{N_1}\sum_{\boldsymbol{x}_i\in R_1(j,s)} y_i \text{ and } \hat{c}_2=\frac{1}{N_2}\sum_{\boldsymbol{x}_i\in R_2(j,s)} y_i.$$

As mentioned before, the stopping criterion for recursive partitioning of the sample, and thus the size to which the tree is allowed to grow, can be pre-specified. Choosing an appropriate size for the tree is important as large trees can result in overfitting, whereas small trees may miss important patterns in the data. The common way to deal with this issue is to choose the size of the tree adaptively, based on the data set at hand, by a process known as cost-complexity

[24] This is where the inclusion of personality traits as categorical instead of continuous variables matters. Splits based on continuous variables are forced to be generated according to a threshold value. Therefore, if a continuous variable is selected for a split, it would not be possible for e.g. the highest and lowest values recorded for that variable to be in the same partition. In the case of categorical variables there is nothing preventing the highest and lowest values being in the same partition.

pruning. A very large tree is grown[25] denoted as T_0, which is almost certain to overfit the data, and then it is pruned to generate a subtree of an appropriate size. Cost-complexity pruning involves introducing a penalty for tree size in the objective function (*RSS* for regression trees) which can be used to tune the depth (complexity) of the tree. In particular, the cost-complexity criterion is defined as:

$$C_a(T) = \sum_{m=1}^{|T|} \sum_{\boldsymbol{x}_i \in R_m(j,s)} (y_i - \hat{c}_m)^2 + \alpha|T|.$$

The criterion is defined for a generic subtree T, where $|T|$ is the number of terminal nodes in the subtree, and $\hat{c}_m = \frac{1}{N_m} \sum_{\boldsymbol{x}_i \in R_m(j,s)} y_i$. The tuning (penalty) parameter α can be used to gauge the size of the tree. Under the objective of minimizing the cost-complexity criterion, setting $\alpha = 0$ results in the original tree grown T_0 as the optimal. A large value for α puts relatively more weight on the parsimony of the tree rather than on the goodness of fit, resulting in a 'small' subtree being optimal, and vice versa. Optimal subtrees are estimated for a range of α values, resulting in a set of optimal subtrees which range in size from $|T_0|$ to the smallest possible subtree of size 1, a subtree with only one node consisting of the entire sample. As James *et al*. (2013) note, for each value of α there is a unique optimal subtree[26].

The size of the penalty, and thus the size of the ultimate subtree, is usually chosen to be the value that minimizes a 10-fold cross-validated sum of squares for the sample at hand (i.e. chosen adaptively)[27].

The regression tree estimation incorporated in the RE-EM tree estimation is implemented through the `rpart` package[28] offered by the statistical software package `R`.

[25] An arbitrary stopping criterion may be used to achieve this. For example, the minimum number of observations in the tree's terminal nodes can be pre-specified. Binary recursive partitioning will be applied on a node as long as the number of observations at that node is above the pre-specified threshold value.

[26] The procedure by which the unique optimal subtree for each value of α is determined is known as weakest link pruning. As the value of α increases, the partition which generates the smallest decrease in *RSS*, and results in terminal nodes for the running subtree is reversed. In this manner, a sequence of nested subtrees is generated, each of which is the optimal subtree, minimizing the cost complexity criterion, for a particular range of α values.

[27] Cross validation is a procedure which determines the value of the tuning parameter based on a pseudo out-of-sample measure of predictive power. In particular, k-fold cross validation involves splitting the sample randomly into k components and then performing the same procedure k times, each time using a different component of the sample as a testing sample (i.e. the data used to determine the out-of-sample predictive power), and the remaining k-1 components as a training sample (i.e. the data used for estimation). The procedure performed in this case is the estimation of the sequence of optimal subtrees for a range of values of the tuning parameter. The value of parameter chosen is the value which generates the subtrees in each sequence such that the cross-validated sum of squares (i.e. sum of squares based on the components used as testing samples) is minimised.

[28] Available on https://cran.r-project.org/web/packages/rpart/rpart.pdf.

4.2 Surrogate variables

Before moving on to the second component of the RE-EM tree estimation procedure, the method by which the issue of observations with missing values can be handled by regression tree estimation will be outlined. Frequently used ways of dealing with missing values for explanatory variables include imputation of the missing values, or ignoring the observations without a complete set of values, which may imply a significant reduction in sample size. Regression tree estimation offers an alternative method for dealing with the issue without having to drop any observations, except for extreme cases[29]. This method exploits the use of the so-called surrogate variables.

As described in subsection 4.1, the choice of the value of a particular variable, known as the primary splitting variable, is made to generate the split of a certain region into two 'smaller' regions. In the case of missing values for explanatory variables, this choice is made based on a splitting criterion which is adjusted to account only for the observations which are not missing the eventual primary splitting variable. Therefore, this implies that a criterion is needed for determining the region to which observations missing the primary splitting variable will go if they are going to be kept in the sample. This is where surrogate variables matter.

A surrogate variable is a variable which aims to mimic the primary splitting variable in terms of the manner in which observations are split into the two 'smaller' regions. This implies that for each surrogate variable an associated value based on which the region can be split in two is required as well. The optimal value for splitting in order to mimic the original split is chosen for every independent variable apart from the primary variable. The surrogate variables are then ranked based on their success which is quantified through a misclassification error associated with each surrogate variable. The misclassification error is defined by the ratio of the number of misclassified observations, when compared to the classification implied in the split generated by the primary variable, over the total number of observations used to derive the misclassification error. Therefore, a relatively low misclassification error places a surrogate variable high in the ranking. For observations missing the primary splitting variable, the path within the regression tree is determined by the combination of the surrogate variable and the associated splitting value with the highest rank. If the value for that variable is missing as well, then it is determined by the surrogate variable with the second highest rank, and so on.

[29] One such example would be the extreme case of an observation with missing values for every explanatory variable included in the estimation.

4.3 Linear mixed effects model

As mentioned before, regression tree estimation constitutes one part of the RE-EM tree estimation. The other part comes in the form of a linear mixed effects model estimation, as outlined by Laird and Ware (1982).

The linear mixed effects model aims to capture the data generating process of a data set which consists of repeated observations of each object (in this case, individuals) in the sample (i.e. a longitudinal or panel data set). For an individual $i \in \{1, \dots, N\}$ at time $t \in \{1, \dots, T_i\}$, the mixed effects model states that:

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\gamma} + \boldsymbol{z}_{it}'\boldsymbol{b}_i + \varepsilon_{it},$$

where

$$\begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT_i} \end{pmatrix} \sim Normal(\mathbf{0}, \boldsymbol{R}_i),$$

and

$$\boldsymbol{b}_i \sim Normal(\mathbf{0}, \boldsymbol{D}).$$

The value of the dependent scalar variable y_{it} is determined by a vector of explanatory variables $\boldsymbol{x}_{it}$ multiplied by a vector of population-level parameters $\boldsymbol{\gamma}$, a vector of explanatory variables $\boldsymbol{z}_{it}$ multiplied by a vector of individual-specific random parameters $\boldsymbol{b}_i$, and a scalar random error component ε_{it}. Population-level parameters are also termed as fixed effects, and individual-specific parameters are termed as random effects, thus constituting a mixed effects model. The linear property of the model is apparent from the determination of y_{it} based on $\boldsymbol{x}_{it}$, and $\boldsymbol{z}_{it}$. In the case that only the intercept is allowed to vary between individuals, z_{it} is a scalar which takes the value of 1.

The effects $\boldsymbol{b}_i$ are assumed to be random, coming from a multivariate normal distribution which is common across individuals, as well as independent across individuals, and uncorrelated with the observed covariates. The error terms are assumed to have a multivariate normal distribution for each individual, and to be independent across individuals. The generic covariance matrix $\boldsymbol{R}_i$ allows for a specification which captures any autocorrelation in the error terms within each

individual, in which case $\boldsymbol{R}_i$ would be non-diagonal[30]. The errors are assumed to be uncorrelated with the effects $\boldsymbol{b}_i$.

Inference for the population-level effects $\boldsymbol{\gamma}$ is based on a maximum likelihood estimator, whereas inference for the individual-specific random effects $\boldsymbol{b}_i$ is based on the work by Harville (1976), who provides an extension of the Gauss-Markov theorem which incorporates the estimation of random effects. Given that $\boldsymbol{y}_i = (y_{i1} \quad \cdots \quad y_{iT_i})'$, $\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}'_{i1} & \cdots & \boldsymbol{x}'_{iT_i} \end{pmatrix}'$, $\boldsymbol{Z}_i = \begin{pmatrix} \boldsymbol{z}'_{i1} & \cdots & \boldsymbol{z}'_{iT_i} \end{pmatrix}'$, and $\boldsymbol{V}_i = \boldsymbol{R}_i + \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}'_i$, Laird and Ware (1982) specify that:

$$\widehat{\boldsymbol{\gamma}} = (\sum_{i=1}^{N} \boldsymbol{X}'_i \boldsymbol{V}_i^{-1} \boldsymbol{X}_i)^{-1} \sum_{i=1}^{N} \boldsymbol{X}'_i \boldsymbol{V}_i^{-1} \boldsymbol{y}_i,$$

and

$$\widehat{\boldsymbol{b}_\iota} = \boldsymbol{D} \boldsymbol{Z}'_i \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\gamma})^{31},$$

where $\widehat{\boldsymbol{\gamma}}$ can be substituted for $\boldsymbol{\gamma}$ in the $\widehat{\boldsymbol{b}_\iota}$ estimator.

However, for both estimators, the parameters which constitute the matrix $\boldsymbol{V}_i$ are unknown. As a result, the matrix can be replaced by an estimator $\widehat{\boldsymbol{V}}_i = \widehat{\boldsymbol{R}}_i + \boldsymbol{Z}_i \widehat{\boldsymbol{D}} \boldsymbol{Z}'_i$. As mentioned by Laird and Ware (1982), both a maximum likelihood (ML) estimator, and a restricted maximum likelihood (RML) estimator are possible candidates for estimating the parameters associated with the covariance matrices. The RML estimator, however, yields an unbiased estimate of the parameters.

Laird and Ware (1982) demonstrate how both the ML and RML estimates can be computed through the expectation-maximization (EM) algorithm. In general, the EM algorithm is an iterative procedure which alternates between an expectation step, which aims at estimating the unobserved components of a model (e.g. some quadratic function of $\boldsymbol{b}_i$ or $(\varepsilon_{i1} \quad \cdots \quad \varepsilon_{iT_i})'$ in this case) given estimates of the model parameters, and a maximization step, which aims at estimating the model parameters given the estimates for the unobserved components. The EM

[30] It is worth noting that in the specification by Laird and Ware (1982) the covariance matrix of the error terms is indexed by *i* such that it represents the dimension of the matrix for a particular individual based on the number of observations recorded for that individual. However, the set of parameters in the covariance matrix does not depend upon i.

[31] From a Bayesian perspective, this estimator can also be seen as the mean of the posterior normal distribution of the random effects conditional on $\boldsymbol{y}_i$. The posterior is obtained by constructing the likelihood component based on the conditional normal distribution of $\boldsymbol{y}_i | \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{b}_i$, and the prior distribution based on the normal distribution of $\boldsymbol{b}_i$ specified above.

algorithm can be shown to always converge with respect to the likelihood function, at least locally.

The linear mixed effects model estimation incorporated in the RE-EM tree estimation is implemented through the nlme package[32] offered by the statistical software package R[33].

4.4 RE-EM tree

The basic model associated with the RE-EM tree is a modification of the linear mixed effects model known as a general mixed effects model. For an individual i at time t:

$$y_{it} = f(\boldsymbol{x}_{it}) + \boldsymbol{z}'_{it}\boldsymbol{b}_i + \varepsilon_{it},$$

where

$$\begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT_i} \end{pmatrix} \sim Normal(\mathbf{0}, \boldsymbol{R}_i),$$

and

$$\boldsymbol{b}_i \sim Normal(\mathbf{0}, \boldsymbol{D}).$$

The linear specification with respect to the observed vector $\boldsymbol{x}_{it}$ is substituted with a generic function f which maps $\boldsymbol{x}_{it}$ to the value of the population-level effect component of the model. Apart from this change, the rest of the model components are specified in the same way as in the linear case above.

RE-EM tree estimation is an iterative procedure which alternates between the estimation of a regression tree, and the estimation of a linear mixed effects model. Given estimates for $\boldsymbol{b}_i$, a regression tree is used to non-parametrically estimate the generic function f, using the modified dependent variable $y_{it} - \boldsymbol{z}'_{it}\widehat{\boldsymbol{b}_\iota}$. Given the estimates for the population-level effects, estimates for the random effects are obtained based on a linear mixed effects model estimation. The estimation and computation procedures for the individual components of this iterative procedure are as specified in subsections 4.1 and 4.3.

Sela and Simonoff (2012) outline the steps to the estimation of the RE-EM tree:

1) Initially set $\widehat{\boldsymbol{b}_\iota}$ to zero.

[32] Available on https://cran.r-project.org/web/packages/nlme/nlme.pdf.
[33] The lme function used to fit the model applies a combination of the Newton-Raphson algorithm and the ECME algorithm, a variant of the EM algorithm which can converge faster than the standard EM algorithm.

2) Iterate between the following steps until convergence[34]:
    a. Regression tree estimation of f using as a dependent variable $y_{it} - \boldsymbol{z}_{it}'\widehat{\boldsymbol{b}_\iota}$. Based on the estimated tree, generate a set of dummy variables representing the terminal nodes of the tree.
    b. Estimate a linear mixed effects model by specifying f to be the set of dummies identified by the regression tree. Obtain $\widehat{\boldsymbol{b}_\iota}$ from the estimation.
3) Based on the estimated coefficients for the set of dummies, replace the predicted dependent variable values for the terminal nodes in order to represent the estimated population-level effect.

The main advantage of the RE-EM tree is that it allows for the flexibility of non-parametric estimation while also accounting for the longitudinal structure of the data. The assumption of additive individual-specific random effects is made based on the general mixed effects model specified above. But the population-level effects are estimated entirely non-parametrically allowing for the flexibility benefits of a regression tree in describing patterns in the data.

The RE-EM estimation is implemented through the REEMtree package[35] offered by the statistical software package R, which uses a combination of the rpart and nlme packages mentioned before.

[34] Convergence is achieved when the change in likelihood or restricted likelihood falls below some threshold value.
[35] Available on https://cran.r-project.org/web/packages/REEMtree/REEMtree.pdf.

5. RESULTS

5.1 RE-EM tree

The estimated RE-EM tree is presented in tabular format in *Table 3*[36]. It characterises each observation as an intersection of conditions based on the explanatory variables used in the estimation of the RE-EM tree[37]. To aid the interpretation of the conditions in the last column of *Table 3*, *Table 2* presents the explanatory variables used along with the labels for each of their values. The rows highlighted in grey in *Table 3* represent the terminal nodes of the RE-EM tree. Therefore, the union of these rows constitutes the original sample used in the estimation. The life satisfaction prediction for the subsample within each node, as presented in *Table 3*, can be used to infer the influence of each individual condition on life satisfaction.

It is also useful to provide a measure which shows the importance of each variable for life satisfaction. This is a feature provided by tree-based estimators in general that may not be directly available in the case of linear regression estimators. Due to the incorporation of explanatory variables in a sequential manner during the construction of a tree, a measure of cumulative *RSS* explained by each individual variable can be composed. *Table 4* presents this measure as a percentage of the total explained *RSS* such that comparisons between variables can be facilitated. As is evident, a significant portion of the explained *RSS* is accounted for by health, followed by the job profile of each individual. Age, and the level of neuroticism also provide substantial contributions relative to the rest of the explanatory variables included.

[36] The RE-EM tree was estimated by using all of the variables available in *Table 2*. For the tree component of the RE-EM tree, the minimum complexity parameter, as used in the `rpart` package, was set to 0.00001. In addition, the concluding choice of the complexity parameter based on the cross-validated sum of squares was made using the one standard error rule. This implies that the eventual parameter chosen was the largest one corresponding to the cross-validated sum of squares that was within one standard error of the minimum (Hastie *et al.*, 2009). For the linear mixed effects component of the RE-EM tree, the intercept was the only element of the vector of individual-specific random effects. In addition, the errors were specified such that there was an autoregressive structure of order 1 within the individuals, and heteroskedasticity across individuals in the sample. The linear mixed effects component was estimated using restricted maximum likelihood.

[37] Tree-based estimators are usually presented in a figure which uses nodes to represent partitions of the original sample, and links between nodes to indicate the path that each observation takes from the root node down, through the tree, to the relevant terminal node. However, in this case the estimated tree is large in the sense that the conventional figure representation would not aid the comprehension of its structure.

Table 2: Variables included in RE-EM tree estimation.

Variable name	Value	Group
Job Status	1	Self-employed
	2	Paid employment
	3	Unemployed
	4	Retired
	5	On maternity leave
	6	Family care
	7	Full-time student
	8	Long-term sick or Disabled
	9	Government training scheme
	10	Unpaid, family business
	11	On apprenticeship
	97	Doing something else
Health	1	Excellent
	2	Very good
	3	Good
	4	Fair
	5	Poor
Marital Status	0	Child under 16
	1	Single
	2	Married
	3	Same-sex civil partnership
	4	Separated
	5	Divorced
	6	Widowed
	7	Separated from civil partner
	8	Former civil partner
	9	Surviving civil partner
	10	Living as couple
Race	1	British, English, Scottish, Welsh, Northern Irish (white)
	2	Irish (white)
	3	Gypsy or Irish traveller (white)
	4	Any other white background
	5	White and black Caribbean (mixed)
	6	White and black African (mixed)
	7	White and Asian (mixed)
	8	Any other mixed background
	9	Indian (Asian or Asian British)
	10	Pakistani (Asian or Asian British)
	11	Bangladeshi (Asian or Asian British)
	12	Chinese (Asian or Asian British)
	13	Any other Asian background
	14	Caribbean (black or black British)
	15	African (black or black British)
	16	Any other black background
	17	Arab
	97	Any other ethnic group
Year	1	2010

	2	2011
	3	2012
	4	2013
	5	2014
	6	2015
	7	2016
	8	2017
	9	2018
Age	15-103	
Log Income	-2.61-12.22	
Country	1	England
	2	Wales
	3	Scotland
	4	Northern Ireland
No. of Children	0-9	
Education	1	Degree
	2	Other higher degree
	3	A-level etc
	4	GCSE etc
	5	Other qualification
	9	No qualification
Sex	1	Male
	2	Female
Personality types	1	Does not apply to me at all
Each ranked on	2	
scale 1 -7	3	
Agreeableness/	4	
Extraversion/	5	
Openness/	6	
Neuroticism/	7	Applies to me perfectly
Conscientiousness		

Table 3: RE-EM tree in tabular format.

Node	Sample size	Life sat. prediction	Path
1	264,518	5.166	Root
2	51,739	4.435	1 ∩ Health = 4, 5
3	11,007	3.592	2 ∩ Job Status = 3, 8, 9, 97
4	5,516	3.212	3 ∩ Health = 5
5	3,072	2.970	4 ∩ Neuroticism > 4
6*	1,957	2.863	5 ∩ Marital Status = 1, 4, 5, 7, 10
7*	1,115	3.246	5 ∩ Marital Status = 2, 3, 6, 8
8*	2,444	3.539	4 ∩ Neuroticism < 5
9	5,491	3.973	3 ∩ Health = 4
10	2,649	3.728	9 ∩ Neuroticism > 4
11*	2,145	3.672	10 ∩ Age < 58
12*	504	4.112	10 ∩ Age > 57
13	2,842	4.202	9 ∩ Neuroticism < 5
14*	2,362	4.135	13 ∩ Age < 60
15*	480	4.606	13 ∩ Age > 59
16	40,732	4.663	2 ∩ Job Status = 1, 2, 4, 5, 6, 7, 10, 11
17	19,987	4.409	16 ∩ Age < 62
18	3,256	3.858	17 ∩ Health = 5
19*	1,390	3.647	18 ∩ Neuroticism > 4
20*	1,866	4.053	18 ∩ Neuroticism < 5
21	16,731	4.516	17 ∩ Health = 4
22	12,164	4.392	21 ∩ Neuroticism > 3
23	4,308	4.175	22 ∩ Marital Status = 1, 4, 5, 6, 7, 8
24*	3,138	4.111	23 ∩ Age > 23
25*	1,170	4.437	23 ∩ Age < 24
26	7,856	4.511	22 ∩ Marital Status = 2, 3, 9, 10
27*	531	4.084	26 ∩ Neuroticism = 7
28	7,325	4.546	26 ∩ Neuroticism = 4, 5, 6
29*	2,730	4.399	28 ∩ Log Income < 7.316
30*	4,595	4.651	28 ∩ Log Income > 7.315
31	4,567	4.845	21 ∩ Neuroticism < 4
32*	2,264	4.721	31 ∩ Log Income < 7.486
33*	2,303	4.983	31 ∩ Log Income > 7.485
34	20,745	4.907	16 ∩ Age > 61
35	5,427	4.401	34 ∩ Health = 5
36*	1,298	4.047	35 ∩ Neuroticism > 4
37*	4,129	4.529	35 ∩ Neuroticism < 5
38	15,318	5.086	34 ∩ Health = 4
39	7,646	4.888	38 ∩ Neuroticism > 3
40	3,466	4.766	39 ∩ Age < 71
41*	732	4.499	40 ∩ Neuroticism > 5
42*	2,734	4.848	40 ∩ Neuroticism = 4, 5
43*	4,180	4.990	39 ∩ Age > 70
44*	7,672	5.279	38 ∩ Neuroticism < 4
45	212,779	5.343	1 ∩ Health = 1, 2, 3

46	77,333	5.122	45 ∩ Health = 3
47	53,376	4.964	46 ∩ Age < 61
48	12,599	4.708	47 ∩ Neuroticism > 4
49	4,570	4.471	48 ∩ Marital Status = 1, 4, 5, 6, 7
50	3,229	4.333	49 ∩ Age > 23
51*	533	3.930	50 ∩ Job Status = 3, 8, 9, 10, 97
52*	2,696	4.455	50 ∩ Job Status = 1, 2, 4, 5, 6, 7, 11
53*	1,341	4.812	49 ∩ Age < 24
54	8,029	4.842	48 ∩ Marital Status = 2, 3, 10
55*	2,568	4.693	54 ∩ Log Income < 7.284
56*	5,461	4.926	54 ∩ Log Income > 7.283
57	40,777	5.043	47 ∩ Neuroticism < 5
58*	3,396	4.637	57 ∩ Job Status = 3, 8, 9
59	37,381	5.080	57 ∩ Job Status = 1, 2, 4, 5, 6, 7, 10, 11, 97
60	17,399	4.988	59 ∩ Log Income < 7.486
61	14, 481	4.943	60 ∩ Neuroticism = 3, 4
62*	1,424	4.721	61 ∩ Marital Status = 0, 4, 5, 6, 7
63*	13,057	4.979	61 ∩ Marital Status = 1, 2, 3, 10
64*	2,918	5.199	60 ∩ Neuroticism < 3
65	19,982	5.161	59 ∩ Log Income > 7.485
66	16,104	5.122	65 ∩ Neuroticism = 3, 4
67	4,718	5.022	66 ∩ Marital Status = 1, 4, 5, 6
68*	2,933	4.884	67 ∩ Age > 22
69*	1,725	5.215	67 ∩ Age < 23
70*	11,386	5.175	66 ∩ Marital Status = 2, 3, 8, 10
71*	3,878	5.318	65 ∩ Neuroticism < 3
72	23,957	5.475	46 ∩ Age > 60
73	10,456	5.325	72 ∩ Neuroticism > 3
74*	4,082	5.174	73 ∩ Year = 3, 4, 5
75	6,374	5.416	73 ∩ Year = 1, 2, 6, 7, 8, 9
76*	1,198	5.147	75 ∩ Job Status = 1, 2, 3, 5, 8, 10, 97
77*	5,176	5.460	75 ∩ Job Status = 4, 6, 7
78	13,501	5.592	72 ∩ Neuroticism < 4
79*	2,766	5.385	78 ∩ Job Status = 1, 2, 3, 6, 8
80*	10,735	5.637	78 ∩ Job Status = 4, 5, 10, 11, 97
81	135,446	5.470	45 ∩ Health < 3
82	100,446	5.385	81 ∩ Job Status = 1, 2, 3, 6, 8, 9, 10, 97
83	46,367	5.261	82 ∩ Neuroticism > 3
84	27,749	5.173	83 ∩ Year = 2, 3, 4, 5
85*	1,618	4.707	84 ∩ Job Status = 3, 8
86	26,131	5.202	84 ∩ Job Status = 1, 2, 6, 9, 10, 97
87	7,471	5.045	85 ∩ Marital Status = 1, 4, 5, 6, 7, 8
88	4,726	4.936	86 ∩ Age > 27
89*	2,103	4.831	87 ∩ Neuroticism > 4
90*	2,623	5.053	87 ∩ Neuroticism = 4
91*	2,745	5.207	86 ∩ Age < 28
92	18,660	5.265	85 ∩ Marital Status = 2, 3, 10
93*	2,302	5.032	92 ∩ Neuroticism > 5

84*	16,358	5.287	92 ∩ Neuroticism = 4, 5
95	18,618	5.393	83 ∩ Year = 1, 6, 7, 8, 9
96	13,724	5.335	95 ∩ Health = 2
97*	4,173	5.195	96 ∩ Marital Status = 1, 4, 5, 6, 7
98*	9,551	5.390	96 ∩ Marital Status = 2, 3, 10
99*	4,894	5.544	95 ∩ Health = 1
100	54,079	5.490	82 ∩ Neuroticism < 4
101	15,945	5.359	100 ∩ Log Income < 7.358
102*	1,791	5.139	101 ∩ Race = 6, 7, 8, 11, 12, 13, 14, 15, 16, 17
103	14,154	5.388	101 ∩ Race = 1, 2, 3, 4, 5, 9, 10, 97
104*	6,476	5.285	102 ∩ Year = 3, 4, 5
105*	7,678	5.463	102 ∩ Year = 1, 2, 6, 7, 8, 9
106	38,134	5.545	100 ∩ Log Income > 7.537
107	24,140	5.491	106 ∩ Health = 2
108*	5,310	5.357	107 ∩ Marital Status = 1, 4, 5
109	18,830	5.527	108 ∩ Marital Status = 2, 3, 6, 10
110*	17,146	5.506	109 ∩ Age < 62
111*	1,684	5.713	109 ∩ Age > 61
112	13,994	5.639	106 ∩ Health = 1
113*	8,885	5.556	112 ∩ Year = 2, 3, 4, 5
114*	5,109	5.755	112 ∩ Year = 1, 6, 7, 8, 9
115	35,000	5.713	81 ∩ Job Status = 4, 5, 7, 11
116	16,559	5.599	115 ∩ Neuroticism > 3
117	11,451	5.527	116 ∩ Health = 2
118*	6,464	5.429	117 ∩ Year = 2, 3, 4, 5
119*	4,987	5.626	117 ∩ Year = 1, 6, 7, 8, 9
120*	5,108	5.747	116 ∩ Health = 1
121	18,441	5.816	115 ∩ Neuroticism < 4
122*	11,193	5.738	121 ∩ Year = 2, 3, 4, 5
123*	7,248	5.905	121 ∩ Year = 1, 6, 7, 8, 9

Notes: The nodes marked with an asterisk and highlighted in grey represent the terminal nodes of the RE-EM tree. The size of the indent for each node in the first column represents the depth of the node in the RE-EM tree structure. Node 1 has no indent given that it is the root node.

Table 4: Variable importance.

Variable	Importance indicator
Health	47
Job status	26
Age	10
Neuroticism	7
Marital status	4
Logarithm of income	2
Education	2
Year of interview	1
Children number	1

Notes: The indicator is constructed based on the percentage of the explained RSS accounted for by each variable throughout the RE-EM tree. The values incorporate splits where a variable is the primary splitting variable, and those where it occurs as a surrogate variable. The total of the values does not add up to 100 as those variables accounting for less than 1% of the explained RSS are not reported.

By using the life satisfaction prediction figures at different nodes, inferences can be made regarding the association of different explanatory variables with life satisfaction. The prediction in this case is made through the estimated population-level effects of the linear mixed effects component of the RE-EM tree.

For example, by looking at nodes 2 and 45 it can be seen that the individuals reporting *Excellent, Very Good,* or *Good* health status have a higher life satisfaction prediction than those reporting *Fair,* or *Poor*. In addition, by looking at nodes 4 and 9, those reporting a *Fair* health status seem to have a higher prediction than those reporting *Poor*. The same observation can be made through nodes 18 and 21. In general, it can be seen that a higher level of health status is associated with a higher life satisfaction prediction throughout the tree. Given the substantial relative importance of health status in *Table 4*, there is evidence that the level of health is paramount in terms of the satisfaction that individuals can derive from their lives. This finding seems to be in agreement with studies such as Gerdtham and Johannesson (2001), and Clark and Oswald (2002) when it comes to the association between well-being and health.

Looking into job status, the major pattern to note is the consistency across the tree in terms of the fact that those reporting to be *Unemployed*, or *Long-term sick or Disabled* are always part of the node which has a lower life satisfaction prediction relative to its pair. This is evident from the pairs of nodes including, but not limited to, nodes 3 and 16, nodes 51 and 52, and nodes 58 and 59. Nodes 3, 51, and 58 represent subsamples which include individuals reporting to be *Unemployed*, or *Long-term sick or Disabled* as opposed to the subsamples represented by nodes 16, 52, and 59 respectively. This is a finding that agrees with the established studies of

Gerdtham and Johannesson (2001), and Clark and Oswald (2002), as well as the newer study of Morrone *et al*. (2019) who use a similar non-parametric estimator.

As far as age is concerned, there seems to be a general pattern of the life satisfaction prediction being lower during the middle part of life, where middle in the above tree is given by the ages between the mid-20s and the 60s. Some examples of the pairs of nodes indicating this include nodes 11 and 12, nodes 14 and 15, and nodes 24 and 25. Node 11 represents a subsample which includes individuals with age less than 58 as opposed to node 12 which represents individuals older than 57 years old who are assigned a higher life satisfaction prediction. Node 14 includes individuals with age less than 60 as opposed to those aged 60 or more represented by node 15 for whom the life satisfaction prediction is higher. Node 24 represents a subsample which includes individuals with age greater than 23 as opposed to node 25 with individuals aged 23 or less having a higher life satisfaction prediction. This finding may not provide a very precise indication of the link between well-being and age, but it is reminiscent of the well-established U-shaped association between the two (Blanchflower and Oswald, 2008) which implies a mid-life nadir in well-being.

Neuroticism is another variable which appears consistently throughout the tree. A higher level of neuroticism is associated with a lower level of predicted life satisfaction. Examples of pairs of nodes showing this include nodes 5 and 8, nodes 10 and 13, and nodes 19 and 20. In particular, nodes 5, 10, and 19 represent subsamples which include individuals with a level of neuroticism recorded by a value of 4 or less as opposed to the subsamples represented by nodes 8, 13, and 20 respectively which record values greater than 4.

The same exercise can be performed for any of the variables that are part of the estimated RE-EM tree. As far as the four most important[38] variables outlined above are concerned, the direction of influence on predicted life satisfaction seems to be approximately consistent throughout the tree. Despite the fact that the influence of individual explanatory variables can be observed, this is something that can be examined using approaches other than trees. The main advantage of tree-based approaches comes in the form of their non-parametric nature. No structure is imposed *a priori*, thus allowing for the tree to be constructed in the manner which is optimal in terms of the goodness of fit in the context of recursive partitioning. This is demonstrated directly by the structure of the estimated RE-EM tree. The recursive binary splitting procedure gives rise to various interactions between explanatory variables. Even

[38] Based on the variable importance summarised by *Table 4.*

though we can look at the influence of individual variables on predicted life satisfaction at different nodes of the tree, this influence is applied only to the observations which exist at that particular node, i.e. the observations which satisfy the conditions on the path which leads to the node under examination.

5.1.1 RE-EM tree insights

As an indication of this type of non-linearity arising in the case of the RE-EM tree, we can consider the early splits that take place during tree construction. The first split is generated based on health status. The individuals reporting *Excellent, Very Good,* or *Good* become part of one node (node 45), and those reporting *Fair,* or *Poor* constitute its pair node (node 2). The subsequent split on each of the two generated nodes occurs based on different splitting variables. For individuals in relatively bad health, the splitting variable which provides the largest *RSS* reduction happens to be job status. In particular, individuals reporting *Unemployed, Long-term sick or Disabled, Government training scheme,* or *Doing something else* move to a node with a lower predicted life satisfaction (node 3) relative to its generated pair node (node 16). For individuals in relatively good health, the splitting variable choice happens to be health status again. In particular, individuals reporting *Excellent,* or *Very Good* move to the node with high predicted life satisfaction (node 81) relative to its generated pair (node 46). Looking deeper into the tree, node 46 is further split based on the age of individuals. Observations with a recorded age variable of 60 or less are associated with lower life satisfaction (node 47) as opposed to those with age greater than 60 (node 72). In addition, node 81, associated with the individuals reporting a relatively high level of health, is further split based on the job status variable into nodes 82 and 115. Even though the same splitting variable as the one applied to node 2 is applied to node 81, the segregation of the values in the support of the variable is different. By looking only at the three early levels of partitioning, we can already observe how the non-linearity manifests itself in the form of different splitting conditions being applied depending on the reported health status associated with each observation. Loosely speaking, depending on health status, the next most important aspect of life associated to reported life satisfaction varies. Going deeper into the estimated tree, several more splitting conditions interact with each other constituting a structure with a significant level of non-linearity.

The structure proposed by the estimated tree is one that can be translated directly to a linear model. This is because the resulting tree is essentially just a set of dummy variables which represent the terminal nodes. The next two subsections focus on estimating the linear model which is composed of this set of dummies through the well-known within estimator, and, using

this estimation, generating predictive margins for various explanatory variables. This is done to facilitate the comparison between the structure offered by the RE-EM tree and the one offered by a classic linear model specification, as well as to aid the interpretation of the tree's complex structure. However, translating the tree's structure into a linear model is not the same as pre-specifying it. Though theoretically possible, estimating the equivalent linear model through more conventional methods would require pre-specifying all of the interactions. To give an idea of how challenging this can be, consider as an example the set of conditions that give rise to the first terminal node presented in *Table 3* (node 6). In order to generate a dummy variable which represents the observations of node 6, the specification of the dummy would have to dictate that the observations for which the value of the variable is 1 (as opposed to 0) satisfy *Poor* health status, *Unemployed, Long-term sick or Disabled, Government training scheme,* or *Doing something else* job status, a neuroticism level strictly greater than 4, and *Single, Separated, Divorced, Separated from civil partner,* or *Living as couple* marital status. Pre-specifying a similar structure, though not impossible, is highly improbable.

Further benefits in the interpretation of the role of individual variables that can be offered by tree-based estimators include the fact that the RE-EM tree estimation allows for the quantification of variable importance, along with the case that only a strict subset of the variables included in the estimation of the RE-EM tree is also involved in the construction of the tree (i.e. variable selection feature). Well-known estimators such as OLS or the within estimator for panel data provide estimated coefficients for every variable included in the linear specification along with collective measures of fit such as the coefficient of determination. In general, however, they do not provide measures of individual variable importance that can facilitate direct comparisons between variables[39], or the variable selection feature. As demonstrated, the variable importance measure can produce a ranking of explanatory variables which can help prioritise life satisfaction associations.

5.2 Within estimator

As mentioned in the previous subsection, the structure of the estimated RE-EM tree can be translated to a linear model. The linear model specification used to represent the estimated tree is such that for individual i at time t:

$$y_{it} = \alpha + \sum_{j=2}^{58} \beta_j z_{jit} + \delta_t + \gamma_i + \varepsilon_{it}.$$

[39] Significance tests do not necessarily provide a direct way of inter-coefficient comparison. Multiple comparisons can be an issue.

y_{it} represents reported life satisfaction. α is a constant term. z_{jit} represents a dummy variable which takes the value of 1 (as opposed to 0) for observations belonging to terminal node j of the estimated tree[40], and β_j is the associated coefficient. δ_t accounts for a time effect which is constant across individuals. γ_i is the unobserved, time-invariant, individual-specific effect. ε_{it} is the typical, normally distributed, random error component. The linear model presented above is estimated through the within estimator, and the results are presented in *Table 5*.

Subsection 5.3 deals with the interpretation of *Table 5*. An interesting application of estimating a linear model which represents the RE-EM tree through the within estimator is the fact that we can draw comparisons to more conventional linear model specifications estimated in the same manner. One such specification can be found in Appendix C. The conventional specification uses all the available explanatory variables in a linear, additive structure without any interactions between them. The particular specification may seem overly simplistic, and thus not a fair comparison. However, loosely speaking, this is the specification that represents the same input that we use in the estimation of the RE-EM tree as well. The fact that the output structures differ substantially between them represents the different ways in which the estimators handle the data, and highlights the contrast between the restrictive nature of parametric approaches and the flexibility of non-parametric approaches. For completion, it should be noted that AIC and BIC can be used for the comparison between the two specifications since they represent a non-nested model comparison. AIC favours the specification based on the RE-EM tree over the conventional linear one, whereas BIC the other way around.

[40] The indexing is arbitrary in this case for the purpose of exposition. It is not related to the way in which nodes are presented in *Table 3* and *Table 5*.

Table 5: Within estimator for life satisfaction based on estimated RE-EM tree.

Variable	Coefficient	Standard error
Year (Default: 2010)		
2011	-0.063***	0.011
2012	-0.120***	0.012
2013	-0.170***	0.012
2014	-0.072***	0.012
2015	-0.016	0.011
2016	-0.022*	0.011
2017	-0.052***	0.014
2018	-0.067	0.036
Terminal node (Default: Node 70)		
6	-1.350***	0.062
7	-1.158***	0.070
8	-1.167***	0.054
11	-0.758***	0.047
12	-0.610***	0.082
14	-0.697***	0.043
15	-0.442***	0.077
19	-0.838***	0.054
20	-0.804***	0.045
24	-0.544***	0.037
25	-0.427***	0.058
27	-0.491***	0.087
29	-0.431***	0.036
30	-0.282***	0.027
32	-0.360***	0.037
33	-0.220***	0.033
36	-0.630***	0.064
37	-0.544***	0.042
41	-0.279***	0.070
42	-0.198***	0.039
43	-0.194***	0.040
44	-0.078*	0.033
51	-0.591***	0.077
52	-0.245***	0.040
53	-0.089	0.055
55	-0.149***	0.037
56	-0.053	0.028
58	-0.311***	0.035
62	-0.204***	0.047
63	-0.078***	0.021
64	-0.102**	0.036
68	-0.150***	0.034
69	0.027	0.046
71	-0.047	0.032
74	0.001	0.033
76	-0.008	0.047
77	0.165***	0.031

79	0.013	0.037
80	0.157***	0.030
85	-0.233***	0.047
89	-0.027	0.044
90	0.026	0.038
91	0.131***	0.036
93	0.027	0.039
94	0.132***	0.021
97	0.108***	0.031
98	0.167***	0.022
99	0.250***	0.028
102	-0.045	0.056
104	0.014	0.027
105	0.087**	0.026
108	0.043	0.029
110	0.116***	0.021
111	0.200***	0.043
113	0.139***	0.026
114	0.248***	0.028
118	0.224***	0.028
119	0.300***	0.028
120	0.356***	0.032
122	0.223***	0.028
123	0.292***	0.029
Constant	5.262***	0.017
Observations	264,518	
Number of individuals	64,260	
AIC	750,745	
BIC	751,469	
Within R-squared	0.023	
Between R-squared	0.241	
Overall R-squared	0.155	

*Notes: Clustered-robust standard errors in parentheses; *p-value < 0.05, **p-value < 0.01, ***p-value < 0.001.The composition of each terminal node is the same as the composition of the equivalent terminal node in Table 3.*

5.3 Predictive margins

The estimated RE-EM tree in *Table 3*, and its counterpart in *Table 5*, provide a substantially dense structure when it comes to the association of various explanatory variables with reported life satisfaction. To make more sense of how individual variables are associated with life satisfaction, the fitted model in *Table 5* is used to generate various predictive margins.

In general, predictive margins represent functions of estimators. In this particular case, the within estimator provides estimates for the coefficients of the linear model specified in subsection 5.2. Predictive margins are calculated for each value in the support of every explanatory variable under consideration. The magnitude of a predictive margin for a fixed value of a particular explanatory variable is just a weighted average of the estimated coefficients. The weights are determined by the amount of observations that exist at each terminal node as a proportion of all the observations which take the fixed value for the explanatory variable under consideration. Given that the predictive margin values are just weighted averages, this implies that standard errors can be calculated using the delta method. As such, for all the figures which follow 95% confidence intervals are also reported.

As mentioned before, and as it can be seen from *Figure 3*, health status appears to have a negative association with life satisfaction. As the reported health status gets worse, the reported life satisfaction drops as well. We can be confident that differences are substantial in terms of life satisfaction between different values of health status because of the observation that the intervals are quite narrow and not overlapping. Furthermore, age seems to demonstrate a mid-life nadir, even if it is on the margin (*Figure 4*). There appears to be more stability after 65 years of age. Another inference based on predictive margins which appears to agree with previous observations made using the estimated RE-EM tree is the case of decreasing life satisfaction as the level of neuroticism increases (*Figure 5*). Again, the confidence intervals are relatively narrow with little overlapping.

An interesting case is the one of the natural logarithm of income (*Figure 6*). First of all, the logarithm of income is considered as a continuous variable. Therefore, the domain of the variable is discretized before calculating the predictive margins. The interesting points for this variable refer, firstly, to the case that up to a value of 10 there seems to be an approximately neutral association between life satisfaction and income, in that no particular direction of impact can be observed. Secondly, for values of the logarithm of income which are greater than

10 there appears to be a positive association with life satisfaction, while the width of the intervals increases.

Lastly, from *Figure 7* we can observe that people who report *Unemployed*, or *Long-term sick or Disabled* for job status appear to be associated with a substantially lower level of life satisfaction. In addition, *Figure 8* shows how *Female* seems to be associated with low life satisfaction relative to *Male*, supported also by non-overlapping confidence intervals. Predictive margins were generated for other explanatory variables as well and can be found in Appendix D.

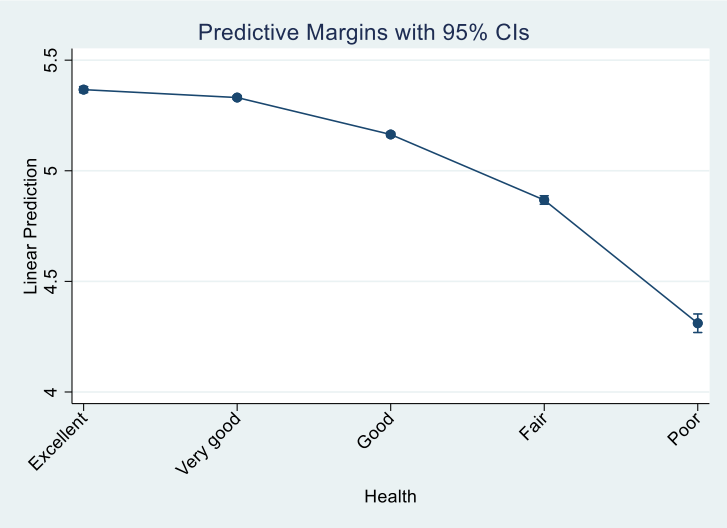

Figure 3: Predictive margins for health with 95% C.I.

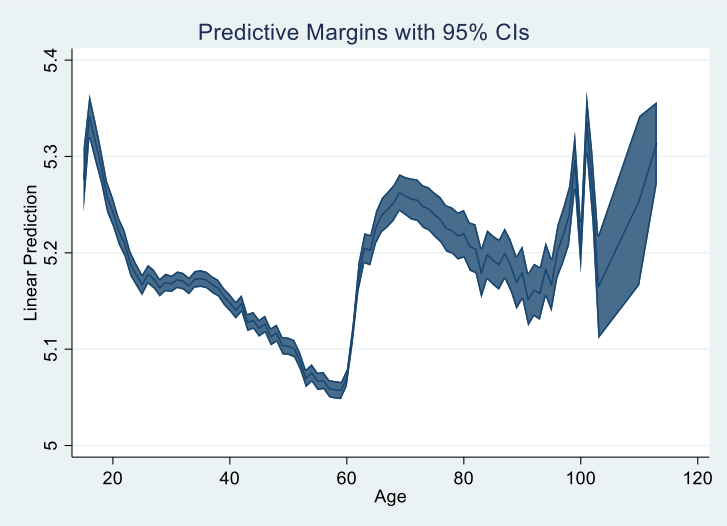

Figure 4: Predictive margins for age with 95% C.I.

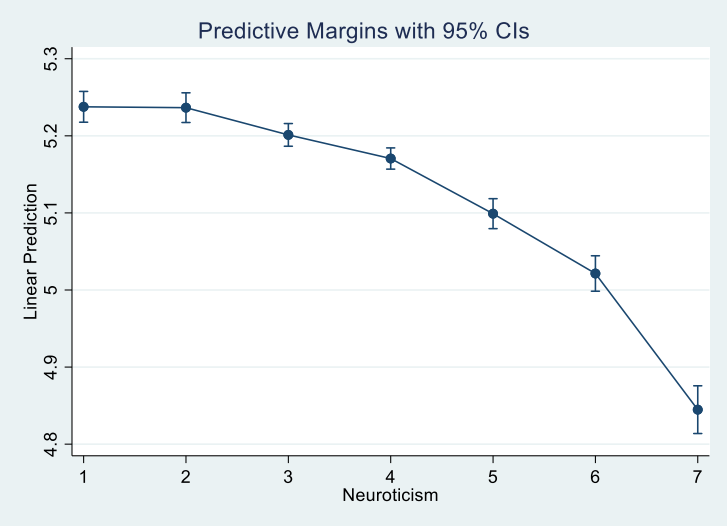

Figure 5: Predictive margins for neuroticism with 95% C.I.

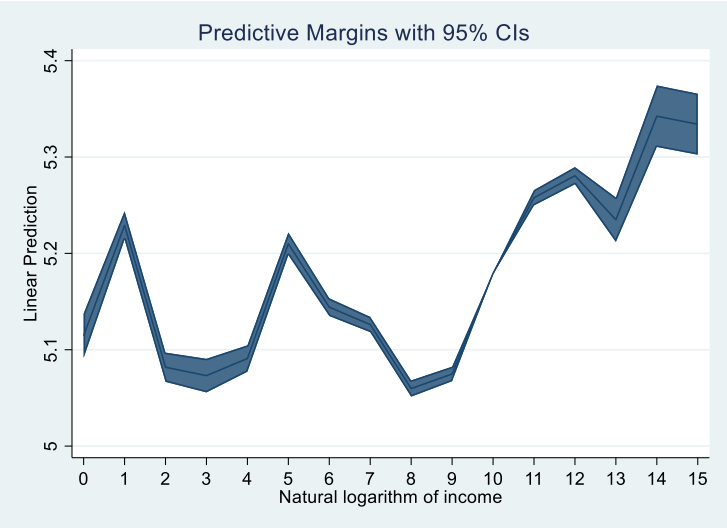

Figure 6: Predictive margins for income with 95% C.I.

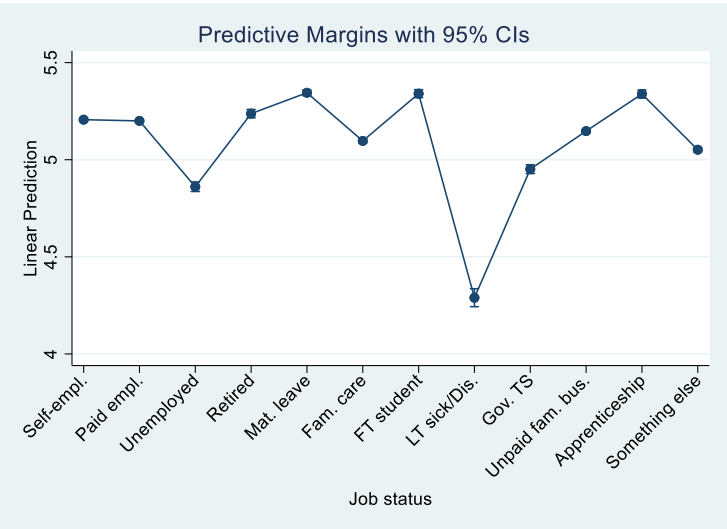

Figure 7: Predictive margins for job status with 95% C.I.

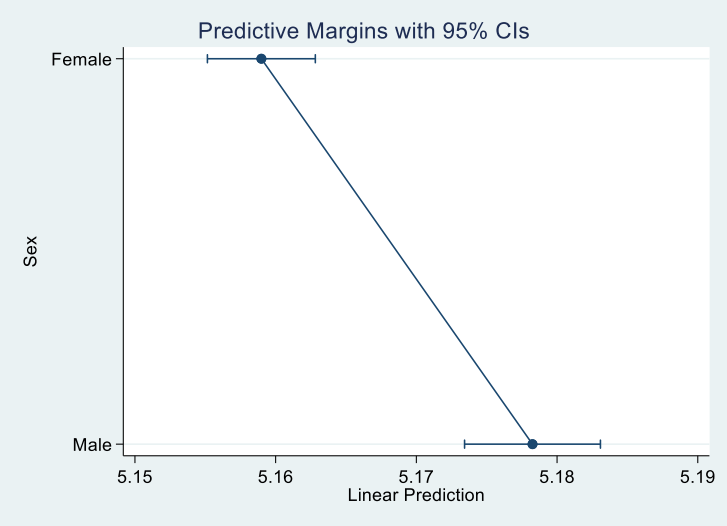

Figure 8: Predictive margins for sex with 95% C.I.

6. CONCLUSION

Understanding which factors influence the determination of well-being, as well as how they might interplay with each other, is of vital importance. Much of the literature has relied on linear techniques such as OLS or the within estimator to identify the variables that are significantly associated with well-being. Valuable insights have emerged. This study aims to contribute to the aforementioned literature by using the RE-EM tree by Sela and Simonoff (2012), a machine learning technique, along with a subjective self-reported life satisfaction measure as a representative of well-being to examine any new insights that might arise by using this alternative methodological approach.

The RE-EM tree is an extension of the famous regression tree method by Breiman *et al*. (1984) which applies to longitudinal data. It is a data mining technique which identifies any patterns in the data used without assuming any particular model structure *a priori*. It offers two major advantages relative to the standard techniques used in the literature. Firstly, it can identify any important non-linearities or interactions between variables which might influence well-being without having to specify them before estimation as in the case of linear models estimated by OLS or the within estimator. Secondly, the RE-EM tree is capable of selecting the most relevant explanatory variables out of a set of variables with an arbitrary size without having to worry about the number of parameters estimated relative to the sample size which might be an issue when using standard techniques.

To aid the interpretation of the estimated RE-EM tree, and improve the comparability of the results with those from the literature which use the standard techniques, a two-step procedure is carried out. Firstly, the structure suggested by the estimated RE-EM tree is used to estimate the equivalent linear model version by the within estimator. Secondly, a set of predictive margins are calculated from the within estimator results which reflect the marginal associations of the various explanatory variables used with the life satisfaction measure. Predictive margins are just functions of the estimated coefficients.

The estimated RE-EM tree proposes a structure with a substantial degree of non-linearity for life satisfaction determination. When the proposed structure is translated into a model which is linear in parameters and estimated by the within estimator, the explanatory power of this model is comparable to the case of directly estimating a linear model with the exact same input as the one supplied to the RE-EM tree estimation procedure. One additional advantage of using the RE-EM tree over the standard methods of analysis is the fact that the importance of the

variables considered in terms of how much explanatory power they contribute to the estimation can be quantified. In doing so, the overwhelming importance of health is observed which amounts for almost half the explanatory power in the estimated tree. It is followed by job status, age, and level of neuroticism of an individual. These variables are well-known determinants of well-being in the literature, but a rank of importance is also provided in the current study.

When using the predictive margins analysis after the application of the within estimator on the RE-EM structure, it can be observed that many of the findings do echo famous literature findings. It is the case that a lower level of overall health is associated with a lower level of life satisfaction. Those who are unemployed are also associated with a lower level of life satisfaction. Furthermore, life satisfaction does exhibit a mid-life nadir as the findings suggest. A higher level of neuroticism is also associated with a lower level of life satisfaction. The aforementioned finding also acts as a testament to how important it is to account for personality traits when considering the analysis of well-being. On a general note, the consistency of findings in this study and across the literature lends confidence in using subjective well-being measures in applied research.

It is interesting to see the application of the RE-EM tree to other well-being concepts outside life satisfaction, such as mental health or positive and negative affect. The revelation of non-linearities is almost certain as it is somewhat inherent in the construction of regression trees. In addition, the variable selection feature of trees allows the simultaneous consideration of every possible aspect that can be considered a candidate explanatory variable with only the most important ones being selected. It is also interesting to see how the structure suggested by the RE-EM estimation, in that people are basically classified into different groups, can offer an easier path towards targeting and implementing group-specific approaches for detecting and alleviating low levels of well-being or mental health. The RE-EM may be offering a simple classification mechanism, yet it still considers the same information as any standard analysis technique. Overall, the findings in the current paper suggest that this type of non-parametric analysis could complement the standard parametric techniques.

7. REFERENCES

Alesina, A., Di Tella, R. and MacCulloch, R., 2004. Inequality and happiness: are Europeans and Americans different?. *Journal of Public Economics*, 88(9), pp.2009-2042.

Becchetti, L., Castriota, S., Corrado, L. and Ricca, E., 2013. Beyond the Joneses: Inter-country income comparisons and happiness. *The Journal of Socio-Economics*, 45, pp.187-195.

Bertrand, M. and Mullainathan, S., 2001. Do People Mean What They Say? Implications for Subjective Survey Data. *American Economic Review*, 91(2), pp.67-72.

Blanchflower, D. and Oswald, A., 2008. Is well-being U-shaped over the life cycle?. *Social Science & Medicine*, 66(8), pp.1733-1749.

Bond, T. and Lang, K., 2019. The Sad Truth about Happiness Scales. *Journal of Political Economy*, 127(4), pp.1629-1640.

Borghans, L., Duckworth, A., Heckman, J. and Weel, B., 2008. The Economics and Psychology of Personality Traits. *Journal of Human Resources*, 43(4), pp.972-1059.

Boyce, C., Brown, G. and Moore, S., 2010. Money and Happiness. *Psychological Science*, 21(4), pp.471-475.

Breiman, L., Friedman, J., Olshen, R. and Stone, C., 1984. *Classification And Regression Trees*. New York: Chapman & Hall.

Clark, A. and Oswald, A., 2002. A simple statistical method for measuring how life events affect happiness. *International Journal of Epidemiology*, 31(6), pp.1139-1144.

Clark, A., 2015. SWB as a Measure of Individual Well-Being. *PSE Working Papers*.

Clark, A., 2018. Four Decades of the Economics of Happiness: Where Next?. *Review of Income and Wealth*, 64(2), pp.245-269.

Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L.J. and Cramer, A.O.J., 2015. State of the art personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, 54, pp. 13–29.

Cox, D.R. and Wermuth, N., 1993. Linear dependencies represented by chain graphs. *Statistical Science*, 8(3), pp.204-218.

Daly, M., Boyce, C. and Wood, A., 2015. A social rank explanation of how money influences health. *Health Psychology*, 34(3), pp.222-230.

Di Tella, R., Haisken-De New, J. and MacCulloch, R., 2010. Happiness adaptation to income and to status in an individual panel. *Journal of Economic Behavior & Organization*, 76(3), pp.834-852.

Diener, E. and Seligman, M., 2004. Beyond Money. *Psychological Science in the Public Interest*, 5(1), pp.1-31.

Easterlin, R., 2005. Building a Better Theory of Well- Being. In: L. Bruni and P. Porta, *Economics and Happiness: Framing the Analysis*. Oxford: Oxford University Press.

Ferrer-i-Carbonell, A. and Frijters, P., 2004. How Important is Methodology for the Estimates of the Determinants of Happiness?. *The Economic Journal*, 114(497), pp.641-659.

Ferrer-i-Carbonell, A. and Van Praag, B., 2008. *Do People Adapt To Changes In Income And Other Circumstances? The Discussion Is Not Finished Yet.*. [online] Pdfs.semanticscholar.org. Available at: https://pdfs.semanticscholar.org/f070/8ec8d40e4925fd9d0cb66a670ca0f0c9095d.pdf.

Ferrer-i-Carbonell, A., 2005. Income and well-being: an empirical analysis of the comparison income effect. *Journal of Public Economics*, 89(5), pp.997-1019.

Ferrer-i-Carbonell, A., 2013. Happiness economics. *SERIEs*, 4(1), pp.35-60.

Frijters, P. and Beatton, T., 2012. The mystery of the U-shaped relationship between happiness and age. *Journal of Economic Behavior & Organization*, 82(2-3), pp.525-542.

Gabriel, S., Mattey, J. and Wascher, W., 2003. Compensating differentials and evolution in the quality-of-life among U.S. states. *Regional Science and Urban Economics*, 33(5), pp.619-649.

Galletta, S., 2016. On the determinants of happiness: a classification and regression tree (CART) approach. *Applied Economics Letters*, 23(2), pp.121-125.

Gerdtham, U. and Johannesson, M., 2001. The relationship between happiness, health, and socio-economic factors: results based on Swedish microdata. *The Journal of Socio-Economics*, 30(6), pp.553-557.

Glenn, N., 2009. Is the apparent U-shape of well-being over the life course a result of inappropriate use of control variables? A commentary on Blanchflower and Oswald (66: 8, 2008, 1733–1749). *Social Science & Medicine*, 69(4), pp.481-485.

Goldberg, L., 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6), pp.1216-1229.

Harville, D., 1976. Extension of the Gauss-Markov Theorem to Include the Estimation of Random Effects. *The Annals of Statistics*, 4(2), pp.384-395.

Hastie, T., Friedman, J. and Tibshirani, R., 2009. *The Elements Of Statistical Learning*. New York: Springer.

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An Introduction To Statistical Learning*.

Konow, J. and Earley, J., 2008. The Hedonistic Paradox: Is homo economicus happier?. *Journal of Public Economics*, 92(1-2), pp.1-33.

Laird, N. and Ware, J., 1982. Random-Effects Models for Longitudinal Data. *Biometrics*, 38(4), p.963.

McCrae, R. and John, O., 1992. An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, 60(2), pp.175-215.

Morrone, A., Piscitelli, A. and D'Ambrosio, A., 2019. How Disadvantages Shape Life Satisfaction: An Alternative Methodological Approach. *Social Indicators Research*, 141(1), pp.477-502.

Mullainathan, S. and Spiess, J., 2017. Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives*, 31(2), pp.87-106.

Oecd-ilibrary.org. 2011. *Divided We Stand*. [online] Available at: https://www.oecd-ilibrary.org/social-issues-migration-health/the-causes-of-growing-inequalities-in-oecd-countries_9789264119536-en.

Oparina, E., Kaiser, C., Gentile, N., Tkatchenko, A., Clark, A. E., De Neve, J.-E., and D'Ambrosio, C., 2022. Human Wellbeing and Machine Learning. [online] Available at: https://arxiv.org/abs/2206.00574.

Osafo Hounkpatin, H., Wood, A., Brown, G. and Dunn, G., 2014. Why does Income Relate to Depressive Symptoms? Testing the Income Rank Hypothesis Longitudinally. *Social Indicators Research*, 124(2), pp.637-655.

Oswald, A. and Powdthavee, N., 2008. Does happiness adapt? A longitudinal study of disability with implications for economists and judges. *Journal of Public Economics*, 92(5), pp.1061-1077.

Oswald, A. and Wu, S., 2010. Objective Confirmation of Subjective Measures of Human Well-Being: Evidence from the U.S.A. *Science*, 327(5965), pp.576-579.

Oswald, A., 2008. On the curvature of the reporting function from objective reality to subjective feelings. *Economics Letters*, 100(3), pp.369-372.

Pfaff, T., 2013. Income Comparisons, Income Adaptation, and Life Satisfaction: How Robust are Estimates from Survey Data?. *SSRN Electronic Journal*.

Proto, E. and Rustichini, A., 2015. Life satisfaction, income and personality. *Journal of Economic Psychology*, 48, pp.17-32.

Roberts, B. and DelVecchio, W., 2000. The rank-order consistency of personality traits from childhood to old age: A quantitative review of longitudinal studies. *Psychological Bulletin*, 126(1), pp.3-25.

Rorer, L., 1965. The great response-style myth. *Psychological Bulletin*, 63(3), pp.129-156.

Schwarz, N. and Clore, G., 1983. Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology*, 45(3), pp.513-523.

Sela, R. and Simonoff, J., 2012. RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine Learning*, 86(2), pp.169-207.

Senik, C., 2004. When information dominates comparison: Learning from Russian subjective panel data. *Journal of Public Economics*, 88(9), pp.2099-2123.

Stiglitz, J., Sen, A. and Fitoussi, J., 2009. [online] Ec.europa.eu. Available at: https://ec.europa.eu/eurostat/documents/118025/118123/Fitoussi+Commission+report.

Tomer, J., 2011. Enduring happiness: Integrating the hedonic and eudaimonic approaches. *The Journal of Socio-Economics*, 40(5), pp.530-537.

Varian, H., 2014. Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2), pp.3-28.

Wood, A., Boyce, C., Moore, S. and Brown, G., 2012. An evolutionary based social rank explanation of why low income predicts mental distress: A 17 year cohort study of 30,000 people. *Journal of Affective Disorders*, 136(3), pp.882-888.

APPENDIX A

Table A.1: Summary statistics of main variables.

Variable	Sample mean	Sample standard deviation
Year		
2010	0.095	0.293
2011	0.154	0.361
2012	0.147	0.354
2013	0.139	0.346
2014	0.137	0.344
2015	0.129	0.336
2016	0.135	0.342
2017	0.058	0.233
2018	0.006	0.076
Job Status		
Self-employed	0.076	0.264
Paid employment	0.483	0.500
Unemployed	0.046	0.210
Retired	0.234	0.423
On maternity leave	0.006	0.075
Family care	0.052	0.222
Full-time student	0.064	0.245
Long-term sick or Disabled	0.033	0.178
Government training scheme	0.001	0.027
Unpaid, family business	0.001	0.025
On apprenticeship	0.001	0.032
Doing something else	0.005	0.068
Health		
Excellent	0.162	0.369
Very good	0.350	0.477
Good	0.292	0.455
Fair	0.142	0.349
Poor	0.054	0.225
Country		
England	0.770	0.421
Wales	0.073	0.260
Scotland	0.092	0.289
Northern Ireland	0.065	0.247
Marital Status		
Child under 16	0.00001	0.004
Single	0.221	0.415
Married	0.520	0.500
Same-sex civil partnership	0.003	0.059
Separated	0.017	0.129
Divorced	0.065	0.247
Widowed	0.058	0.234
Separated from civil partner	0.0002	0.015
Former civil partner	0.0001	0.008
Surviving civil partner	0.0001	0.009

Living as couple	0.115	0.319
No. of Children	0.488	0.910
Education		
Degree	0.244	0.429
Other higher degree	0.120	0.325
A-level etc	0.215	0.411
GCSE etc	0.209	0.407
Other qualification	0.092	0.290
No qualification	0.120	0.324
Natural logarithm of Income	7.418	0.655
Age	48.350	18.284
Personality		
Agreeableness	5.634	1.033
Extraversion	4.591	1.302
Openness	4.552	1.302
Neuroticism	3.561	1.437
Conscientiousness	5.483	1.103
Sex	1.557	0.497
Life Satisfaction	5.168	1.490
Race		
British, English, Scottish, Welsh, Northern Irish	0.815	0.388
Irish	0.021	0.144
Gypsy or Irish traveller	0.0002	0.014
Any other white background	0.027	0.162
White and black Caribbean	0.006	0.080
White and black African	0.002	0.048
White and Asian	0.004	0.062
Any other mixed background	0.003	0.059
Indian	0.030	0.170
Pakistani	0.023	0.151
Bangladeshi	0.012	0.110
Chinese	0.004	0.063
Any other Asian background	0.009	0.096
Caribbean	0.016	0.126
African	0.018	0.133
Any other black background	0.001	0.038
Arab	0.003	0.056
Any other ethnic group	0.003	0.059

APPENDIX B

Table B.1: Partial correlation table.

	Life sat.	Rude	Thorough job	Talkative	Worries lot	Original	Forgiving	Lazy
Life satisfaction	1.000	-0.012	0.090	0.020	-0.141	-0.017	0.019	-0.001
Rude	-	1.000	0.037	0.053	0.046	0.070	-0.107	0.181
Does a thorough job	-	-	1.000	0.151	0.031	0.079	0.021	-0.091
Talkative	-	-	-	1.000	0.113	0.073	0.016	0.028
Worries a lot	-	-	-	-	1.000	0.053	0.040	0.021
Original	-	-	-	-	-	1.000	0.069	0.029
Forgiving nature	-	-	-	-	-	-	1.000	-0.018
Lazy	-	-	-	-	-	-	-	1.000
Sociable	0.076	-	-	-	-	-	-	-
Nervous	0.005	0.010	-	-	-	-	-	-
Artistic	0.018	0.00001	0.062	-	-	-	-	-
Kind	-0.006	-0.225	0.020	0.064	-	-	-	-
Efficient	0.024	0.017	0.289	0.017	0.023	-	-	-
Reserved	-0.004	0.014	0.050	-0.215	0.066	-0.012	-	-
Relaxed	0.106	-0.010	0.023	-0.001	-0.314	0.050	0.098	-
Active Imagination	-0.007	0.047	-0.015	0.072	0.002	0.314	-0.001	0.040
	Sociable	**Nervous**	**Artistic**	**Kind**	**Efficient**	**Reserved**	**Relaxed**	**Active Imag.**
Life satisfaction	-	-	-	-	-	-	-	-
Rude	-0.014	-	-	-	-	-	-	-
Does a thorough job	-0.017	-0.023	-	-	-	-	-	-
Talkative	0.358	-0.007	-0.042	-	-	-	-	-
Worries a lot	-0.034	0.393	0.001	0.061	-	-	-	-
Original	0.071	-0.063	0.182	-0.051	0.088	-	-	-
Forgiving nature	0.071	0.038	0.034	0.279	-0.015	0.032	-	-
Lazy	0.013	0.134	0.064	-0.010	-0.184	0.071	0.028	-
Sociable	1.000	-0.010	0.053	0.108	0.081	-0.179	0.103	0.060
Nervous	-	1.000	0.053	0.075	-0.011	0.203	-0.185	0.005
Artistic	-	-	1.000	0.084	-0.009	-0.007	-0.018	0.225
Kind	-	-	-	1.000	0.329	0.043	0.041	0.047
Efficient	-	-	-	-	1.000	0.123	0.104	0.055
Reserved	-	-	-	-	-	1.000	0.138	0.037
Relaxed	-	-	-	-	-	-	1.000	0.141
Active Imagination	-	-	-	-	-	-	-	1.000

Notes: The partial correlation table is generated by using 40,068 observations from the third wave of Understanding Society. Personality characteristics are captured only in the third wave. Only observations for which there are no missing values in any of the variables used to generate the partial correlation network are used.

APPENDIX C

Table C.1: Within estimator based on conventional linear model.

Variable	Coefficient	Standard error
Year (Default: 2010)		
2011	-0.092***	0.011
2012	-0.157***	0.011
2013	-0.204***	0.011
2014	-0.103***	0.011
2015	-0.004	0.011
2016	-0.008	0.011
2017	-0.036*	0.014
2018	-0.050	0.037
Job Status (Default: Self-employed)		
Paid employment	-0.010	0.018
Unemployed	-0.245***	0.025
Retired	0.119***	0.024
On maternity leave	0.161***	0.038
Family care	-0.029	0.026
Full-time student	0.147***	0.029
Long-term sick or Disabled	-0.396***	0.035
Government training scheme	-0.166	0.134
Unpaid, family business	-0.025	0.102
On apprenticeship	0.087	0.087
Doing something else	-0.027	0.046
Health (Default: Excellent)		
Very good	-0.084***	0.010
Good	-0.252***	0.012
Fair	-0.536***	0.015
Poor	-0.995***	0.024
Country (Default: England)		
Wales	0.136	0.086
Scotland	-0.018	0.100
Northern Ireland	0.244	0.186
Marital Status (Default: Child under 16)		
Single	-0.702	1.258
Married	-0.631	1.258
Same-sex civil partnership	-0.491	1.260
Separated	-0.845	1.259
Divorced	-0.702	1.259
Widowed	-0.804	1.259
Separated from civil partner	-0.993	1.279
Former civil partner	-0.406	1.338
Surviving civil partner	-0.791	1.328
Living as couple	-0.587	1.258
No. of Children (Default: 0)		
1	0.030	0.016
2	0.002	0.021
3	0.024	0.033
4	-0.052	0.058

5	-0.064	0.126
6	0.047	0.219
7	-0.183	0.419
8	0.236	0.293
9	-0.376***	0.024
Education (Default: Degree)		
Other higher degree	-0.036	0.046
A-level etc	0.059	0.034
GCSE etc	0.100*	0.043
Other qualification	0.069	0.074
No qualification	0.075	0.072
Natural logarithm of Income	0.042***	0.007
Constant	5.755***	1.260
Observations	264,518	
Number of individuals	64,260	
AIC	750,758	
BIC	751,282	
Within R-squared	0.023	
Between R-squared	0.204	
Overall R-squared	0.134	

*Notes: Clustered-robust standard errors in parentheses; *p-value < 0.05, **p-value < 0.01, ***p-value < 0.001.*

APPENDIX D

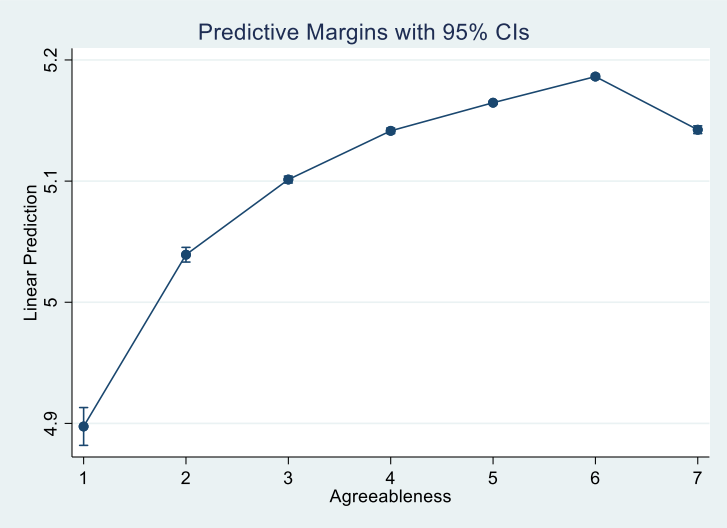

Figure D.1: Predictive margins for agreeableness with 95% C.I.

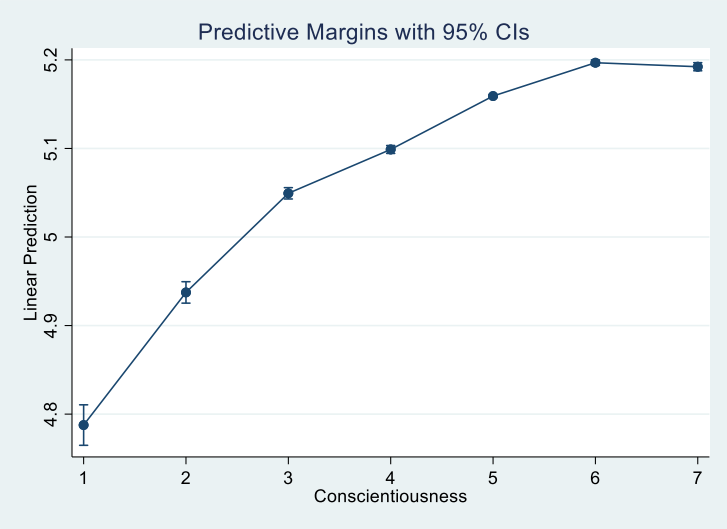

Figure D.2: Predictive margins for conscientiousness with 95% C.I.

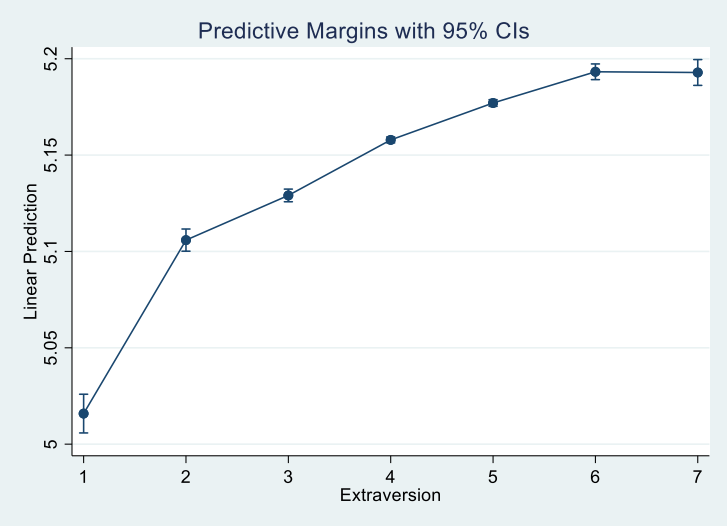

Figure D.3: Predictive margins for extraversion with 95% C.I.

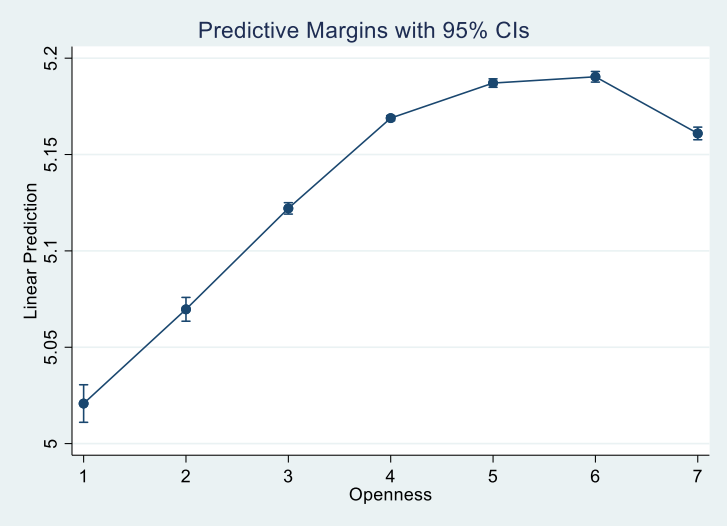

Figure D.4: Predictive margins for openness with 95% C.I.

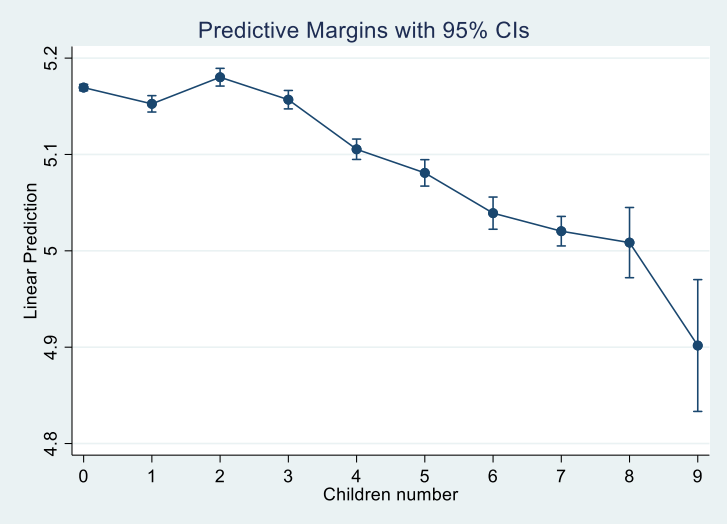

Figure D.5: Predictive margins for number of children with 95% C.I.

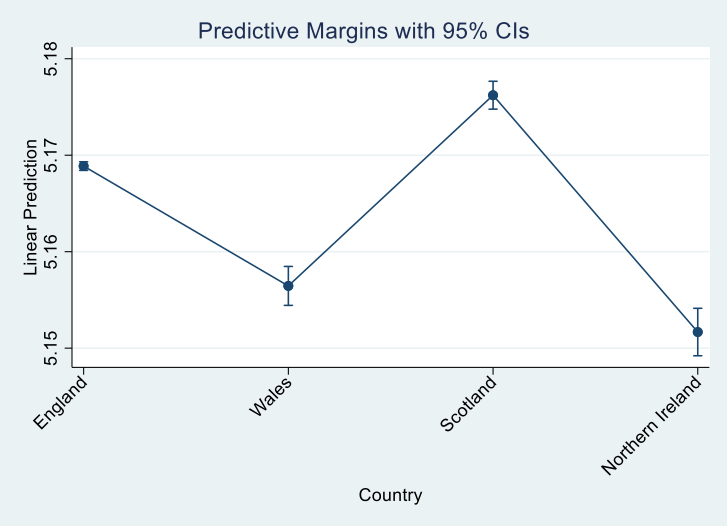

Figure D.6: Predictive margins for country with 95% C.I.

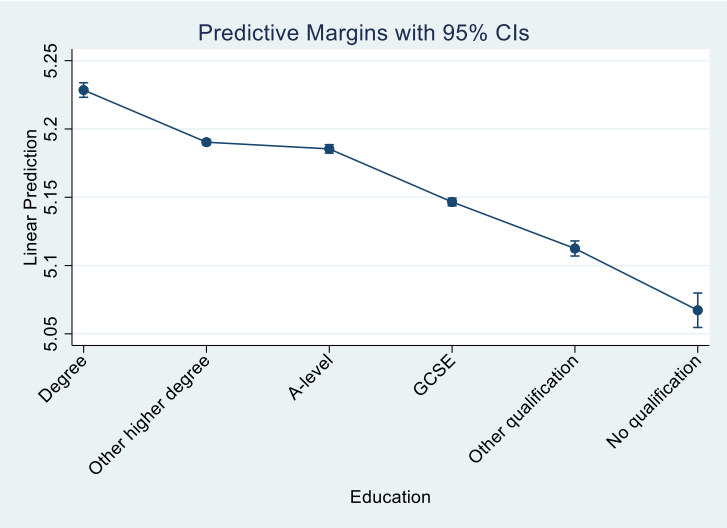

Figure D.7: Predictive margins for level of education with 95% C.I.

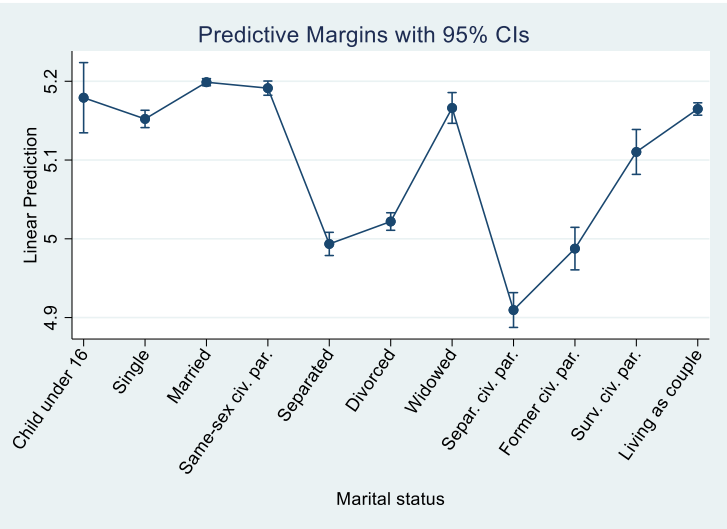

Figure D.8: Predictive margins for marital status with 95% C.I.

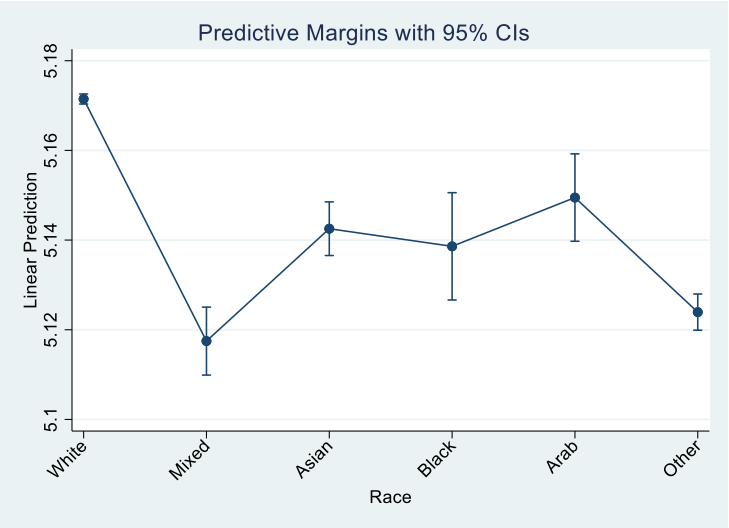

Figure D.9: Predictive margins for race with 95% C.I.

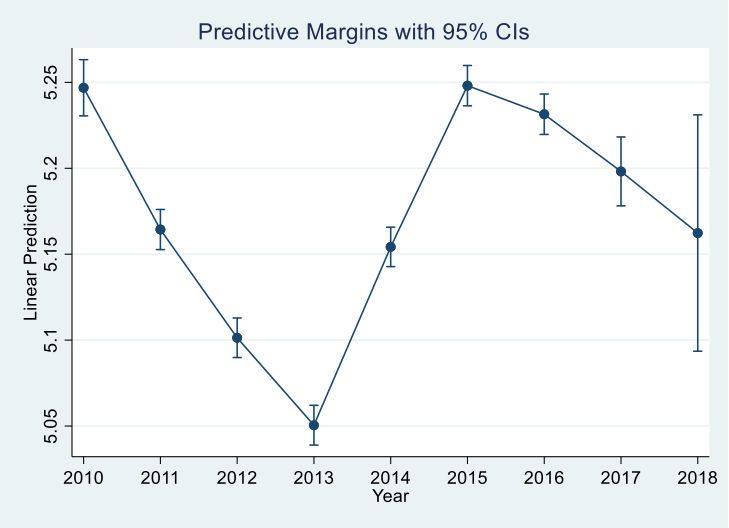

Figure D.10: Predictive margins for year with 95% C.I.

CHAPTER 3: THE IMPACT OF THE COVID-19 PANDEMIC ON THE DETERMINATION OF WELL-BEING

Abstract: 2020 saw a major deterioration in the UK's average level of mental health as a result of the COVID-19 virus outbreak and the associated policies implemented, specifically lockdowns, in an attempt to limit the impact of the pandemic. Many life aspects considered as major determinants of well-being and mental health have been adversely affected in the period since. Major societal shocks may have also influenced the direction and magnitude of the associations of these aspects with well-being or mental health. This study investigates the possibility of a structural break in the determination equation for mental health. Recognizing changes in well-being determination during crisis periods is important if interventions are to target the anticipated deterioration in well-being. The potential of a structural break is examined for males and females separately with respect to aspects such as cohabitation, whether or not there is a child in the household, the employment status, the frequency of loneliness feelings, the health status, the hours worked per week, and absolute income. The structural break is examined in the context of a known break which coincides with the implementation of the first lockdown in the UK (March 2020), and significant changes in the coefficients of the mental health determination equation are detected. We also find tentative evidence of a second structural break during the summer of 2020, shortly after many of the UK's COVID-19 restrictions had been eased.

1. INTRODUCTION

The COVID-19 virus was identified in Wuhan, China, during December 2019. On the 11th of March 2020 the World Health Organisation declared a pandemic, and on the 26th March 2020 the first national lockdown was imposed in the UK, which was followed by two more national lockdowns on the 5th November 2020 and the 6th January 2021[1]. The COVID-19 pandemic represented a time of unprecedented challenges for the UK and the rest of the world, impacting every aspect of life. For many individuals it led to concerns about their health, social restrictions, money worries and job insecurity, all of which led to a large deterioration in mental health or well-being[2] (see, Banks and Xu, 2020; Chandola *et al*., 2020; Daly *et al*., 2020; Banks *et al*., 2021). However, although the pandemic is now receding, given its severity, this raises the question as to whether it has led to changes in some of the determinants of mental health. We investigate the extent to which this is the case in what follows.

Understanding the determinants of mental health is of great importance to policymakers. Mental health is a key component of subjective well-being in its own right and is also a risk factor for future physical health and longevity. Recognizing how the determinants of mental health change during a crisis enables policy makers to target interventions at those in most need, mitigating deteriorations in mental health. Although crises are often accompanied by structural breaks to other sectors of the economy, to the best of our knowledge no study has investigated whether a crisis can also lead to a structural break in the determinants of mental health. Indeed, changes in market forces brought about by the 2008 global financial crisis led to a structural break in the housing market (see, Martins *et al*., 2021) and oil market (see, Fan and Xu, 2011), while the 1997 Asian financial crisis led to a structural break in the real estate market (see, Gerlach *et al*., 2006) and the stock market (see, Baek and Jun, 2011).

Using data from the UK's Household Longitudinal Survey and Understanding Society's COVID-19 web survey, this paper investigates whether there has been a structural break in the determinants of mental health during the pandemic that coincides with the start of the first lockdown in the UK.

[1] A brief timeline of events regarding COVID-19 in the UK can be found on https://www.instituteforgovernment.org.uk/sites/default/files/timeline-lockdown-web.pdf.
[2] Well-being and mental health are assumed to be equivalent in the context of the present study.

However, given the nature of the pandemic, assuming this is the only structural break may be naïve. Therefore, as part of the robustness checks we investigate the presence of a second structural break with an unknown date.

In the analysis that follows, we focus on the effect seven key variables had on mental health during the pandemic: partnership status, feelings of loneliness, the presence of children, health status, employment status, hours worked, and absolute income. These variables are often considered to be the main determinants of well-being (see, Ferrer-i-Carbonell, 2013; Clark, 2018).

We find evidence of two structural breaks in the determinants of mental health. The first occurs at the start of the pandemic. The results show that the pandemic and the first lockdown in the UK had a significant detrimental effect on mental health. There were also some noticeable changes in some of the key variables that are thought to affect mental health, particularly the role loneliness plays in determining mental health. The results suggest that part of the increase in mental distress during the pandemic, and the accompanying social isolation, is explained by changes in feelings of loneliness and/or changes in the association of feelings of loneliness with mental health. The social isolation element of the pandemic also disturbed family life, leading to increased mental distress for those with school-aged children in the household after the temporary halt of the educational system. The pandemic, being a public health emergency, also disturbed the relationship between physical health and mental health. One of the most consistent findings in the well-being literature before the pandemic is that good health carries a well-being premium. This is significantly reduced during the pandemic. The rising concern about physical health seems to have taken a toll on the mental health of those who never had to genuinely worry about it.

Another worrying finding was the effect the disruptions to the labour market had on mental health. Prior to the pandemic there was a premium in mental health associated with having a job relative to being unemployed. Now we observe that for some employed individuals their mental health became similar to those without a job, while engaging in more hours of work is now linked with elevated mental health.. Amongst this disturbance, we find that the UK government seems to have done a decent job in guarding the mental health of individuals participating in the furlough scheme, while the work-from-home routine established at the start of the pandemic seems to favour those who do not always work from home.

We also find tentative evidence of a change in the role income plays in determining mental health. Prior to the pandemic, it was often argued that the social comparison of income was more important in determining mental health than the absolute level of income (see, Boyce *et al*., 2010; Becchetti *et al*., 2013). During the pandemic absolute income, which reflects the standard of living, is still an important factor for mental health. However, we find differences in social comparisons between males and females which might reflect inherent characteristics in terms of how each gender views others during crises. Males appear to care more about the general income level of their peers, and not necessarily how they rank in their reference group, whereas females care more about where they rank among their peers, regardless of their level of income.

Investigating some of these effects further as part of the robustness checks we find tentative evidence of a second structural break. This occurs during the summer of 2020, shortly after many of the UK's COVID-19 restrictions had been eased. It is mainly driven by the reduced mental health burden of those experiencing heightened feelings of loneliness. None of the other variables seem to restore their pre-pandemic association with mental health. An interesting subject for future research is the examination of the mental health determinants to find out if they shift back to their pre-pandemic state as more data becomes available with time, or if there is some form of permanent structural change. Any form of permanent change would mean that policy makers targeting interventions which mitigate mental distress should not revert back to their pre-pandemic tactics after the Covid-19 crisis recedes.

The remainder of the paper is organised as follows. Section 2 presents an overview of the literature. Section 3 introduces the data. Section 4 outlines the econometric model and results for known and unknown structural break dates. The last section concludes the paper.

2. LITERATURE REVIEW

2.1 Mental health and COVID-19

The COVID-19 pandemic has spurred a large part of the recent literature on well-being and mental health. Many of the studies focus on the impact that different changes in the lives of individuals, brought about by the coronavirus, have had on well-being. First and foremost, COVID-19 is fundamentally a public health issue. The pandemic has also restricted social interaction as citizens were asked to remain in isolation for extended periods of time in order to limit the contagious behavior of COVID-19, on more than one occasion and for many countries. The fact that people were not able to attend their jobs, and businesses had to face reduced demand and operational difficulties due to the unprecedented situation present only some of the immediate economic implications of the pandemic. These, and several other less obvious facets of life during the coronavirus period, can prove to be threatening for the well-being and mental health of individuals.

The initial negative impact of COVID-19 on mental health is evident across many studies (see, Banks and Xu, 2020; Chandola *et al.*, 2020; Daly *et al*., 2020; Banks *et al.*, 2021). Banks and Xu (2020) suggested that the overall population effect of the pandemic on mental health during April 2020 was approximately the same in magnitude as the pre-pandemic difference in mental health between the top and bottom income quintiles in the UK. During the same period, Chandola *et al.* (2020) found that 29% of the UK adults classified as not having a common mental disorder (CMD)[3] less than a year earlier could be considered to have a CMD in April 2020. Daly *et al*. (2020) noted that this elevated proportion of individuals with mental health issues persisted in the UK in May and June 2020[4]. However, they suggest that there is evidence of recovery in the months following April 2020. This finding is supported by Banks *et al*. (2021), and Chandola *et al.* (2020). Daly and Robinson (2021) suggested that by September 2020 distress levels among UK adults were similar to pre-pandemic levels. On the other hand, Quintana-Domeque and Proto (2022) using UK data from March 2021 argued that mental health levels did not revert to their pre-pandemic state.

Many studies find that the initial effects of COVID-19 in the UK were most prominent among women, young individuals, and ethnic minorities (Banks *et al*., 2021). Banks and Xu (2020) find evidence of a stronger negative impact for women and young adults in the first two months of the UK lockdown; Daly *et al*. (2020) using data after April 2020 find that problems were

[3] As measured by responses to the 12-item General Health Questionnaire (GHQ-12).
[4] Their definition of a mental health issue agrees with the CMD definition used by Chandola *et al.* (2020).

highest among females and individuals aged 18-34 years; Li and Wang (2020) using the notion of general psychiatric disorders echoed the aforementioned findings; Pierce *et al*. (2020) and Niedzwiedz *et al*. (2021b) find worsening of the GHQ-12 score among young adults and females during April 2020.

Studies which examined how the trajectory of psychological distress varied among UK individuals used latent class mixture modelling to classify individuals into four trajectories of distress, which include continuously low, temporarily elevated, repeatedly elevated, and continuously elevated. Again, it was found the young and females had the highest risk of belonging to any of the latter three groups (Ellwardt and Präg, 2021). Pierce *et al*. (2021) identified GHQ-12 trajectories using data from April to October 2020. They found individuals with pre-existing mental health issues, and individuals belonging to an ethnic minority group had a higher probability of being classified into trajectories that exhibited a worsening of mental health at the start of the pandemic, which was sustained through the COVID-19 period; or a trajectory characterised by a small deterioration at the start of the pandemic but a sustained decline in mental health across the time period considered.

Given the consistent finding that COVID-19 had a relatively larger deteriorating impact on the mental health of young people, restricting the analysis of the impact on the sample of young people can provide an interesting perspective. In a study which compared the impact of the pandemic on the mental health of young people between the UK and China, Liu *et al.* (2021) suggested that characteristics like gender, loneliness, nationality, and psychotherapy[5] had significant predictive power in terms of the GHQ-12 score among individuals with a mean age of 23 in both the UK and China.

In addition to these broad categories, other studies also based on UK data examined more specific characteristics. Chandola *et al.* (2020) suggested that, by July 2020, self-reported loneliness, and financial stressors, such as unemployment, remained significant determinants of the probability that an individual experienced a common mental disorder. Daly *et al.* (2020) also pointed to high-income and educated individuals to be among those groups of people who experienced the largest increase in mental health problems during April 2020. Li and Wang (2020) mentioned living with a partner and having a job as protective factors against general

[5] Represented by a binary variable capturing the demand for psychotherapy.

psychiatric disorders during the start of the pandemic, while experiencing COVID-19 symptoms increased the likelihood of facing deteriorated mental health.

In the context of mental health trajectories, as presented by Ellwardt and Präg (2021), the authors found that long-term distress was evident for those living without a partner, individuals without a job, experiencing a reduction in income, with pre-existing health issues, and experiencing COVID-19 symptoms. Niedzwiedz *et al.* (2021a) showed that psychological distress was more common amongst individuals reporting COVID-19 symptoms in relation to individuals without probable COVID-19 infection, which held for up to seven months after the report of the symptoms. In their trajectory analysis for mental health, Pierce *et al.* (2021) also noted that pre-existing health conditions, and whether the individual was living in a deprived neighbourhood, reinforced the probability that an individual was part of a group identified as having poor mental health throughout the pandemic period.

A substantial component of the pandemic's influence on the every-day lives of individuals was through the major shock to the labour market. Crossley *et al.* (2021) found from April to May 2020 that approximately half of the individuals in their sample faced at least a 10% reduction in household earnings relative to the pre-pandemic period. Another major change came in the form of reduced working hours. Ferry *et al.* (2021) found that 42% of employees reported a reduction in working hours by April 2020. Despite not finding evidence of an association between reduced working hours and mental health, individuals facing reduced work hours due to a permanent lay-off, or those with reduced hours due to caring responsibilities, experienced a deteriorating impact on mental health. Giovanis and Ozdamar (2021) studied the influence of the transition towards working from home on mental well-being. They found that a shift towards working 'always from home' had a negative impact on mental well-being, whereas shifting towards working 'from home on occasion' did not make a difference.

The UK government reacted rapidly to reducing the impact of COVID-19 on the job market through introducing furloughing and the job retention scheme. Chandola *et al.* (2020) suggested that this might have had a positive impact on mental health, as they found that the probability that an individual experienced a common mental disorder was similar among furloughed employees and those whose jobs were not affected in the early months of the pandemic.

Social restrictions and isolation have also been identified as important features of the pandemic resulting in feelings of loneliness. Bu *et al.* (2020) found those aged 18-30, individuals living

on their own, and those with low household income were the most likely to experience evaluated feelings of loneliness. Li and Wang (2020) found young adults, females, individuals without a job, or without a partner living with them were also groups characterised by an elevated risk of loneliness, together with those who experienced COVID-19 symptoms.

2.2 The relative income hypothesis

This paper is also concerned with how income and the social comparisons of income are associated with well-being. Duesenberry (1949) proposed that the mechanism determining the impact of income on well-being involves individuals evaluating their own income against a reference point. His hypothesis suggested that this reference point is based on the income of others who are 'relevant' to one's self, termed as an individual's reference group. Easterlin (1974) provided empirical evidence conformable with the hypothesis. He presented data suggesting happiness increased with income both among and within countries at a point in time, and yet remained constant over time despite the elevated income associated with economic growth. This means that absolute income is unlikely to be the sole income-related determinant of well-being. Instead it can be explained by also including an adjustable reference point when determining wellbeing.

A common assumption is that the reference group is determined by observable characteristics such as age, location, and the level of education. Boes *et al*. (2010), Boyce *et al*. (2010), and Ferrer-i-Carbonell (2005) used this approach. Another important aspect is whether individuals evaluate their position in the reference group ordinally or cardinally. For example, a purely ordinal evaluation of income can suggest that aggregate welfare is not affected by income inequality since individuals only care about their ranking in terms of income. Someone will be ranked first and someone else last in any possible scenario[6]. This is not the case for income-related arguments with a cardinal nature[7].

Two specifications are investigated in the current study, namely reference income and rank of income. Reference income is typically defined as the arithmetic mean of the reference group's income distribution (see, Ferrer-i-Carbonell, 2005; Becchetti *et al*., 2013; Clark and Oswald, 1996; and Senik, 2004). Ferrer-i-Carbonell (2005), Becchetti *et al*. (2013), and Clark and Oswald (1996) found a negative relationship of reference income with happiness, life satisfaction and job satisfaction, respectively. In contrast, Senik (2004) demonstrated that

[6] The marginal impact of income on well-being can still vary according to the income inequality level.
[7] Cardinal in this study refers to some form that is not a pure ordinal rank.

reference income can have a positive influence on life satisfaction since it can incorporate information for potential improvements in the economic condition of individuals.

An alternative approach to incorporate social comparison is the Decision-by-Sampling theory proposed by Stewart *et al*. (2006) which uses a pure ordinal rank. Boyce *et al.* (2010) suggested that a pure ordinal rank-based specification outperforms the specification which uses the absolute and relative income combination. This is suggested through a significance comparison of the estimated coefficients in the case in which they are simultaneously included in the specification. This potential dominance is evident in other areas of research including income satisfaction and job satisfaction (see, Boes *et al.*, 2010; Card *et al.*, 2012), mental distress (see, Wood *et al.*, 2012), mental health (see, Daly *et al.*, 2015), and depressive symptoms (see, Osafo Hounkpatin *et al.*, 2015). However, there are studies like Clark *et al*. (2009) modelling satisfaction with economic conditions who support the coexistence of reference income and the ordinal component in the same specification.

The Decision-by-Sampling component represents the normalized ranking of each individual within the reference group to which they belong. It is calculated as follows for individual i with absolute income x_i in reference set R_i:

$$f_i = \frac{k_i - 1}{n_i - 1} \in [0,1], \quad (1)$$

where $k_i = |\{x \in R_i | x \leq x_i\}|$ is the ordinal rank of income within the reference group distribution and n_i is the size of the reference group.

3. DATA

3.1 UK Household Longitudinal Study and the COVID-19 Web Survey

This chapter uses data from the UK Household Longitudinal Study (UKHLS; Understanding Society), an ongoing panel survey of more than 40,000 households that began in 2009 and Understanding Society's COVID-19 web survey.

The COVID-19 web survey is a monthly survey that was introduced to record the impact of the COVID-19 pandemic. Individuals aged 16 and over participating in the main UKHLS survey (waves 8 or 9) were invited to complete web-surveys from April 2020 to September 2021[8]. We combine waves 1 to 10 of UKHLS which were collected between 2009 and 2020 with the COVID-19 surveys to create a longitudinal data set.

Both surveys ask respondents questions on a range of socioeconomic issues such as health, labour market activity, income, and family-life, as well as welfare and psychological status. Some variables (such as education and retirement) are not available in the COVID-19 survey, and instead their value immediately prior to the start of the pandemic is assumed. Other questions are not asked in the same way in both surveys and hence a harmonization process is used to create a set of comparable variables. All variables used are defined in Appendix A.

The sample used in the analysis is an unbalanced panel data set of 204,301 observations from 17,456 individuals, 7,343 males and 10,113 females. The pre-COVID-19 sample incorporates only individuals who participate in the COVID-19 version of the survey. The selection is such that all observations included have no unobserved values for any of the variables incorporated in the analysis[9;10].

It should be noted that although the COVID-19 data is only available over a relatively short period of time, there is sufficient variation in the data for panel data techniques to be used in estimation. See Appendix B where we present the within sample standard deviations for the variables for which the structural break is investigated.

[8] Recorded during April, May, June, July, September, and November 2020, and January, March, and September 2021. The surveys were conducted in the last week of each month.

[9] The criterion for participating in at least one wave of the COVID-19 study precedes the one for having no missing values for any of the variables incorporated in the analysis. As such, there can still be individuals with observations only in the pre-COVID-19 period in the estimation sample. Given that the within estimator is used, two observations per individual are also required implicitly. Thus observations from individuals appearing only once in the combined data set are not included in the estimations.

[10] Since the aim of this paper is not to estimate the population regression function, but rather to generate an approximation of the well-being determination process, sampling weights are not employed. See Angrist, and Pischke (2009) for a more detailed discussion of the use of survey weights.

3.2 General Health Questionnaire (GHQ-12)

Mental health is measured using the 12-item General Health Questionnaire (GHQ-12). It assesses the severity of non-specific mental distress over the past 2 weeks using a 4-point scale (from 0 to 3). It covers problems such as difficulties with concentration, sleep, decision-making, strain, and feeling overwhelmed. The score in each dimension is used to generate a total score ranging from 0 to 36, with higher scores indicating better mental health.

This GHQ-12 is frequently used in the literature to capture well-being, mostly from the perspective of mental health or psychological distress (see, Clark and Oswald, 2002; Wood *et al*., 2012; and Brown *et al*., 2015). Goldberg *et al*. (1997) provided early support for the robustness of the 12-item GHQ measure in comparison with more complex indicators when studying psychological disorders. Further details of the measure are in Appendix C.

Figure 1 plots the average GHQ per month for males and females, respectively from 2009 to 2021.

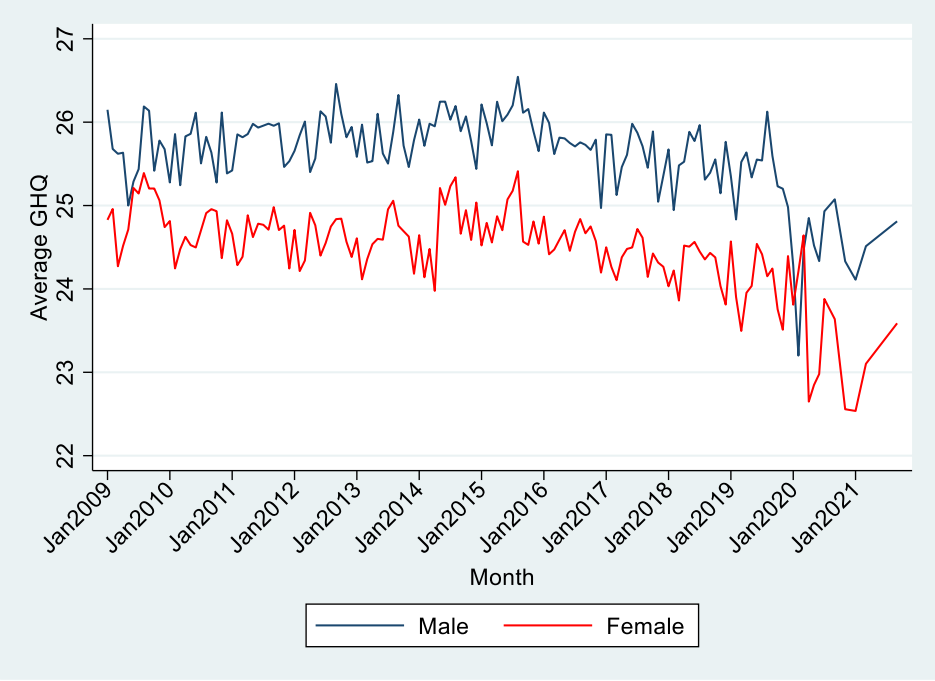

Figure 1: Average GHQ per month 2009-2021.

The score is relatively stable up to January 2020, with females having a consistently lower GHQ score. However, from January 2020 onwards, at which time the COVID-19 virus' arrival in the UK can be tracked, a decrease in the average GHQ score is observed. Using March 2020 as a cut-off, the month when the first lockdown was introduced in the UK, the average GHQ score is 23.71 for the period after the threshold (24.63 for males; 23.06 for females), as

compared to 25.06 for the period up to and including this cut-off (25.75 for males; 24.57 for females). This represents deterioration in mental health of more than 1 point on average on the GHQ 36-point scale.

4. TESTING FOR A STRUCTURAL BREAK

4.1 Econometric specification

The literature on structural breaks often assumes that the timing of a break is a nuisance parameter that should be estimated along with the rest of the parameters in the model. However, if the timing of a break is assumed to be known *a priori* the model is no different from the ordinary linear one (Wang, 2015). Therefore, as a first approach to evaluating the impact of COVID-19 on mental health, a single break is assumed *a priori* with the timing coinciding with the announcement and imposition of the first lockdown in the UK.

Specifically we estimate a regression of the form:

$$GHQ_{it} = \alpha + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \boldsymbol{z}'_{it}\boldsymbol{\gamma_1} + 1(t > c)\boldsymbol{z}'_{it}\boldsymbol{\gamma_2} + d_t + h_i + \varepsilon_{it}, \quad (2)$$

where GHQ_{it} measures mental health for individual i at time t. $\boldsymbol{z}_{it}$ is a vector of explanatory variables for which the impact on mental health changes depending on the timing t of the observation. If $t \leq c$ then the impact on mental health is captured by the vector of coefficients $\boldsymbol{\gamma_1}$, otherwise it is captured by $\boldsymbol{\gamma_1} + \boldsymbol{\gamma_2}$. The parameter c captures the timing of the structural break and $1(\cdot)$ is the indicator function. The vector $\boldsymbol{x}_{it}$ represents a set of explanatory variables for which the impact on mental health remains constant over time, captured by the vector of coefficients $\boldsymbol{\beta}$. The scalar d_t captures a time effect which is fixed across individuals, and the scalar h_i captures individual-specific heterogeneity which is constant across time. Finally, ε_{it} is the random error component, which is uncorrelated across individuals, but can be correlated across observations for the same individual.

Controlling for individual fixed effects enables us to capture key personality traits and other non-cognitive aspects of behaviour such as motivation, drive and ambition that might influence responses to questions on mental health, and at the same time may be associated with other explanatory variables such as income or feelings of loneliness. As such, the standard way to approach estimation is the within group estimator which removes the requirement that the unobserved terms be uncorrelated with the explanatory variables of interest by demeaning each variable with the across-time average of each individual, thus removing any components which are time-invariant as captured by h_i[11].

[11] Clustered-robust standard errors at the individual level are used in estimation in an attempt to avoid inefficiency of the estimator. Inefficiency can result from ignoring the possible serial correlation at the individual level which may imply errors smaller in size than what they should be. An inclusive list of all the assumptions required for each panel data method to achieve consistency and efficiency is given by Cameron and Trivedi (2005).

In specifying a linear regression, we are making the assumption that mental health is a cardinal rather than an ordinal construct. This is not a concern as studies dealing with this issue suggest that the distinction of ordinality and cardinality for self-reported measures is relatively inconsequential for the results (see, Ferrer-i-Carbonell and Frijters, 2004; and Pfaff, 2013).

4.2 COVID-19 structural break

The results are displayed in *Table 1* for men and women, respectively. In column 1 we present the results without the structural break, while the results with the structural break are presented in columns 2 to 4. The COVID-19 interaction effects are used to test for the presence of a structural break, and represent the additional effects of the covariates on mental health during the pandemic. *Table 2* presents the aggregate effect each variable has on mental health during the pandemic and is the sum of the pre-pandemic effect and the structural break coefficient, i.e. $\boldsymbol{\gamma_1 + \gamma_2}$.

In line with the summary statistics, column 1 (*Table 1*) shows that for both males and females the time dummies associated with the period after the introduction of the first lockdown are positive and significant, indicating that mental health declined significantly from April 2020 onwards, relative to the base year of 2017. Given that the regression controls for the main determinants of mental health, this acts as an indication that the existing set of control variables have a different impact on mental health after the onset of the pandemic[12].

Indeed, after allowing for a structural break, column 2 shows that for both males and females the coefficients on the time dummies during the pandemic are now either lower in magnitude or are insignificant. This can be taken as an indication that the introduction of the structural break allows the model to account for the deterioration in mental health during the pandemic through channels other than the time dummies. Another indication is the lower AIC and BIC values for the model with a structural break. The AIC and BIC are likelihood-based measures of how well the model fits the data, where lower values indicate a better fit. Furthermore, performing a Wald test for the joint significance of the coefficients representing the structural break in column 2 generates a p-value of 0.000 for both males and females, causing us to reject the null hypothesis of no structural break in the determinants of mental health after the onset of the pandemic.

[12] In order to incorporate any seasonality in well-being, column 1 is re-estimated where dummy variables for both year and month are included in place of the existing time period variable. The results show significantly positive coefficients on the dummy variables for 2020 and 2021 with reference year still being 2017. Results available on request.

We turn now to examine the effect the remaining variables have on mental health and how this changes during the pandemic. *Table 1* (column 2) shows that for both males and females although not living with a partner had no effect on mental health before the pandemic, it has a detrimental effect during the pandemic. For males the overall effect is also negative and significant (*Table 2*), while for females it remains insignificant (with a p-value for the total effect of 0.109).

What we are observing may, however, be capturing the effect of loneliness, which had a large role to play during the pandemic. We investigate the extent to which this is the case in column 3 where we control for loneliness. As already mentioned, since this variable is only available from 2017 onwards, the pre-pandemic period is now somewhat reduced. To understand how this affects the inferences made, the original model which excludes loneliness is also re-estimated using this reduced sample (column 4) for comparison purposes.

Now we find that although not living with a partner has a positive effect on mental health before the pandemic (males only), its effect is largely insignificant during the pandemic. In contrast, loneliness has a detrimental effect on mental health for both males and females, the effect of which is increased during the pandemic. This suggests that part of the positive impact living with a partner has on mental health in a model which does not control for loneliness is because living with a partner averts some feelings of loneliness. This finding echoes those of Chandola *et al*. (2020) and Liu *et al*. (2021) who suggest that self-reported loneliness has significant predictive ability in terms of mental distress during the pandemic.

It should be noted that the AIC and BIC values of the estimation which incorporates loneliness (*Table* 1: column 3) are substantially lower than the one which does not (*Table* 1: column 4), indicating the explanatory power added to the model by the loneliness variable. This suggests that part of the increase in mental distress during the pandemic is explained by changes in feelings of loneliness and/or changes in the association of feelings of loneliness with mental health. However, comparisons between the results using the shorter panels in columns 3 and 4 show that, apart from cohabitation status, omitting loneliness has no substantial impact on the inferences made for the rest of the variables. Therefore, in the remainder of our discussion of *Tables 1* and *2* we proceed by considering only the results from column 2.

In terms of the remaining variables, column 2 shows that for both males and females there is increased mental distress associated with having a child under 15 in the household during the pandemic period. This might arise due to the increased stress that home schooling and a lack

of social interaction had on family life, especially during the initial lockdown in March 2020 when restrictions were at their tightest. One major difference between males and females is the negative effect having a child has on mental health for males in the pre-COVID-19 period, an effect which is not present for females. This may arise because of the different roles and responsibilities males and females have in the household, with women often being more responsible for childcare than men.

For both men and women, having a long-term illness also has a negative effect on mental health, although this effect is much less pronounced during the pandemic. In fact, *Table 2* shows that the overall effect of having a long-term illness on mental health is no longer significant during the pandemic. This finding points towards a significant change in the perception or behaviour towards illness during the pandemic. During the pandemic, those with no underlying health conditions, perhaps for the first time, had a reason to be concerned about their health, while those with a long-standing health condition may already have been experiencing a heightened level of mental distress because of their health[13]. It should be noted that this result remains in a specification which also incorporates a dummy variable capturing whether or not individuals experience any symptoms of COVID-19 in the period before each web-survey (see, *Table D.1* in Appendix D). This variable in itself has a significantly negative coefficient, indicating an increased level of mental distress associated with experiencing COVID-19 symptoms.

In terms of employment, in the pre-pandemic period having a job or being retired has a positive effect on mental health, while hours worked has a detrimental effect. During COVID-19, the positive effect of having a job or being retired is somewhat reduced, making the overall effect insignificant for those in employment. This might arise because of the increased uncertainty in the job market due to COVID-19, while for those who are retired it may represent the increased risk from COVID-19 due to age. In contrast, hours worked has a positive effect on mental health during the pandemic, perhaps also reflecting concerns about job security. What we are observing may also reflect the importance employment had in combating feelings of social isolation that arose during the pandemic. As Ferry *et al*. (2021) note, the extent to which reduced working hours can have a dent on mental health depends on the source of the reduction.

[13] Another possible explanation for the effect long-term illness had on mental health during the pandemic might arise due to the relief some individuals felt at discovering they were no longer at risk of developing serious complications if they were to contact COVID-19. As we learn more about COVID-19 some conditions are no longer thought of as putting individuals at serious risk of developing serious complications.

In particular, permanent layoffs or an increased burden of caring responsibilities can result in a reduction in the level of mental health.

Finally, for both males and females there is a positive effect of absolute income on mental health before the pandemic, but no additional effect is detected during the pandemic. We return to the effect income has on mental health later in the paper when we look at the impact of social comparisons of income.

Taken together these findings highlight the significant impact the pandemic had on mental health, as well as on the determinants of mental health, particularly those related to feelings of loneliness.

Table 1: Within estimator with no and one structural break for years 2009-2021.

	1		2		3		4	
Variable	**Males**	**Females**	**Males**	**Females**	**Males**	**Females**	**Males**	**Females**
Time period identifier (Default: 2017)								
2009	0.163	0.430***	0.147	0.400***	-	-	-	-
	(0.103)	(0.0975)	(0.103)	(0.0978)	-	-	-	-
2010	0.0596	0.282***	0.0389	0.250***	-	-	-	-
	(0.0780)	(0.0732)	(0.0783)	(0.0735)	-	-	-	-
2011	0.201***	0.258***	0.185**	0.229***	-	-	-	-
	(0.0726)	(0.0713)	(0.0727)	(0.0715)	-	-	-	-
2012	0.308***	0.262***	0.293***	0.237***	-	-	-	-
	(0.0706)	(0.0688)	(0.0707)	(0.0689)	-	-	-	-
2013	0.159**	0.206***	0.145**	0.185***	-	-	-	-
	(0.0711)	(0.0685)	(0.0711)	(0.0686)	-	-	-	-
2014	0.333***	0.355***	0.318***	0.334***	-	-	-	-
	(0.0661)	(0.0658)	(0.0662)	(0.0659)	-	-	-	-
2015	0.428***	0.352***	0.421***	0.341***	-	-	-	-
	(0.0623)	(0.0638)	(0.0622)	(0.0638)	-	-	-	-
2016	0.182***	0.207***	0.177***	0.201***	-	-	-	-
	(0.0590)	(0.0583)	(0.0590)	(0.0584)	-	-	-	-
2018	-0.0549	-0.0982*	-0.0467	-0.0890	0.0134	0.0159	0.0559	-0.0328
	(0.0583)	(0.0574)	(0.0584)	(0.0574)	(0.0727)	(0.0697)	(0.0757)	(0.0732)
2019	-0.141*	-0.160**	-0.136	-0.151*	-0.0554	-0.0443	0.0137	-0.111
	(0.0831)	(0.0781)	(0.0831)	(0.0781)	(0.107)	(0.100)	(0.112)	(0.106)
January 2020	-0.240	-0.257	-0.225	-0.273	-0.511	-0.162	-0.285	-0.338
	(0.447)	(0.425)	(0.446)	(0.424)	(0.501)	(0.446)	(0.520)	(0.479)
February 2020	-1.063**	-0.437	-1.085**	-0.446	-0.748	-0.536	-0.766	-0.521
	(0.499)	(0.531)	(0.499)	(0.534)	(0.612)	(0.666)	(0.633)	(0.661)
March 2020	-0.0234	0.205	-0.0476	0.222	-1.278*	0.331	-1.038	0.149
	(0.730)	(1.442)	(0.733)	(1.419)	(0.697)	(1.402)	(0.642)	(1.614)
April 2020	-0.768***	-1.796***	-0.560*	-1.405***	-0.476	-0.807***	-0.803**	-1.278***
	(0.0820)	(0.0822)	(0.300)	(0.267)	(0.335)	(0.287)	(0.358)	(0.308)
May 2020	-1.008***	-1.528***	-0.800***	-1.144***	-0.766**	-0.624**	-1.078***	-1.021***
	(0.0866)	(0.0842)	(0.307)	(0.272)	(0.341)	(0.291)	(0.364)	(0.311)
June 2020	-1.169***	-1.473***	-0.993***	-1.111***	-1.003***	-0.719**	-1.299***	-1.019***
	(0.0893)	(0.0863)	(0.310)	(0.274)	(0.341)	(0.293)	(0.365)	(0.314)
July 2020	-0.639***	-0.649***	-0.474	-0.292	-0.476	0.00770	-0.771**	-0.211
	(0.0874)	(0.0847)	(0.310)	(0.275)	(0.342)	(0.294)	(0.366)	(0.315)
September 2020	-0.562***	-0.861***	-0.421	-0.543*	-0.372	-0.236	-0.756**	-0.485
	(0.0867)	(0.0885)	(0.312)	(0.277)	(0.342)	(0.296)	(0.367)	(0.316)
November 2020	-1.307***	-1.915***	-1.162***	-1.593***	-1.017***	-1.044***	-1.476***	-1.549***
	(0.0949)	(0.0930)	(0.313)	(0.276)	(0.344)	(0.295)	(0.367)	(0.317)
January 2021	-1.474***	-1.936***	-1.330***	-1.619***	-1.049***	-0.879***	-1.621***	-1.558***
	(0.0999)	(0.0956)	(0.317)	(0.278)	(0.347)	(0.294)	(0.371)	(0.317)
March 2021	-1.015***	-1.268***	-0.878***	-0.954***	-0.781**	-0.461	-1.187***	-0.898***
	(0.0971)	(0.0944)	(0.315)	(0.276)	(0.344)	(0.294)	(0.368)	(0.316)
September 2021	-0.652***	-0.835***	-0.519	-0.530*	-0.530	-0.296	-0.814**	-0.467
	(0.0973)	(0.0921)	(0.319)	(0.278)	(0.349)	(0.297)	(0.373)	(0.318)
Not living with a partner (Default: No)								
Yes	-0.325***	-0.0684	-0.131	0.0122	0.388**	0.238	0.271	0.111
	(0.114)	(0.0961)	(0.119)	(0.1000)	(0.192)	(0.164)	(0.210)	(0.178)
Yes (during COVID-19)	-	-	-0.600***	-0.218*	-0.156	0.0189	-0.450***	-0.210
	-	-	(0.147)	(0.114)	(0.158)	(0.122)	(0.172)	(0.129)
Child under 15 in the household (Default: No)								
Yes	-0.405***	-0.184**	-0.253***	-0.0292	-0.123	0.265*	-0.147	0.180
	(0.0810)	(0.0744)	(0.0860)	(0.0790)	(0.164)	(0.143)	(0.179)	(0.154)

Yes (during COVID-19)	-	-	-0.431***	-0.506***	-0.118	-0.303***	-0.129	-0.296**
	-	-	(0.120)	(0.113)	(0.131)	(0.116)	(0.142)	(0.128)
How often feels lonely (Default: Hardly ever or never)								
Some of the time	-	-	-	-	-1.566***	-2.080***	-	-
	-	-	-	-	(0.121)	(0.0980)	-	-
Often	-	-	-	-	-5.276***	-5.267***	-	-
	-	-	-	-	(0.330)	(0.210)	-	-
Some of the time (during COVID-19)	-	-	-	-	-1.100***	-0.847***	-	-
	-	-	-	-	(0.135)	(0.110)	-	-
Often (during COVID-19)	-	-	-	-	-1.965***	-2.453***	-	-
	-	-	-	-	(0.400)	(0.253)	-	-
Long-standing health condition (Default: No)								
Yes	-0.193**	-0.188***	-0.322***	-0.358***	-0.237**	-0.0423	-0.332***	-0.0239
	(0.0754)	(0.0711)	(0.0807)	(0.0781)	(0.118)	(0.113)	(0.128)	(0.123)
Yes (during COVID-19)	-	-	0.344***	0.425***	0.287**	0.304***	0.310**	0.293**
	-	-	(0.113)	(0.108)	(0.122)	(0.113)	(0.135)	(0.121)
Employed (Default: No)								
Yes	0.767***	0.396***	1.156***	0.756***	1.362***	0.860***	1.415***	0.852***
	(0.125)	(0.0928)	(0.140)	(0.103)	(0.262)	(0.188)	(0.285)	(0.205)
Retired	1.576***	1.366***	1.733***	1.575***	1.635***	1.680***	1.532***	1.766***
	(0.132)	(0.114)	(0.137)	(0.116)	(0.288)	(0.240)	(0.302)	(0.257)
Yes (during COVID-19)	-	-	-0.815***	-0.884***	-1.268***	-0.623***	-1.164***	-0.682***
	-	-	(0.255)	(0.186)	(0.308)	(0.214)	(0.332)	(0.233)
Retired (during COVID-19)	-	-	-0.589**	-0.890***	-1.074***	-1.428***	-0.801**	-1.248***
	-	-	(0.260)	(0.240)	(0.294)	(0.254)	(0.316)	(0.273)
Hours worked per week	0.00140	-0.000278	-0.00478**	-0.0067***	-0.00103	-0.00110	-0.00119	0.00385
	(0.00186)	(0.00192)	(0.00220)	(0.00227)	(0.00398)	(0.00378)	(0.00435)	(0.00411)
Hours worked per week (during COVID-19)	-	-	0.0142***	0.0152***	0.0122***	0.00727*	0.0145***	0.00907**
	-	-	(0.00333)	(0.00331)	(0.00426)	(0.00394)	(0.00463)	(0.00425)
Absolute income	0.0622***	0.0615***	0.0520***	0.0555***	-0.00577	0.0480*	-0.00849	0.0368
	(0.0125)	(0.0119)	(0.0141)	(0.0132)	(0.0257)	(0.0253)	(0.0275)	(0.0271)
Absolute income (during COVID-19)	-	-	0.0349	0.0193	0.118***	0.0237	0.115***	0.0222
	-	-	(0.0292)	(0.0279)	(0.0314)	(0.0312)	(0.0336)	(0.0336)
Housing tenure (Default: Owned)								
Owned (mortgage)	-0.368***	-0.339***	-0.339***	-0.302***	0.0263	0.0532	0.0248	0.0195
	(0.0813)	(0.0791)	(0.0810)	(0.0788)	(0.126)	(0.125)	(0.137)	(0.135)
Rented	-0.332**	-0.0701	-0.285**	-0.0351	0.190	0.228	0.226	0.276
	(0.140)	(0.122)	(0.139)	(0.121)	(0.248)	(0.215)	(0.262)	(0.233)
Other	-0.352	-0.152	-0.279	-0.129	0.00377	0.0524	-0.212	0.335
	(0.331)	(0.307)	(0.329)	(0.308)	(0.370)	(0.348)	(0.421)	(0.380)
Government Office Region (Default: North East)								
North West	1.375**	1.310*	1.442**	1.289*	-1.412	1.692	-0.950	1.754
	(0.678)	(0.739)	(0.674)	(0.745)	(2.212)	(1.141)	(2.652)	(1.411)
Yorkshire and The Humber	1.038*	0.538	1.109**	0.538	-0.419	1.553	0.228	1.467
	(0.544)	(0.775)	(0.539)	(0.781)	(1.979)	(1.011)	(2.311)	(1.241)
East Midlands	0.829	0.606	0.881	0.600	-0.565	0.519	0.0668	0.457
	(0.589)	(0.730)	(0.585)	(0.734)	(1.860)	(1.219)	(2.212)	(1.469)
West Midlands	1.608**	0.710	1.661**	0.712	-1.827	1.753	-0.984	2.030
	(0.699)	(0.761)	(0.697)	(0.766)	(2.116)	(1.225)	(2.420)	(1.515)
East of England	1.070**	0.746	1.132**	0.738	-1.441	0.888	-0.318	1.404
	(0.533)	(0.748)	(0.529)	(0.750)	(1.778)	(1.363)	(2.118)	(1.611)

London	0.481	0.793	0.570	0.766	-1.564	0.607	-1.369	1.033
	(0.564)	(0.738)	(0.559)	(0.741)	(1.816)	(1.193)	(2.175)	(1.461)
South East	1.469***	0.837	1.540***	0.842	-1.012	1.381	-0.323	1.550
	(0.557)	(0.735)	(0.553)	(0.738)	(1.808)	(1.256)	(2.186)	(1.528)
South West	1.376**	0.749	1.468**	0.747	-0.911	1.418	-0.223	1.572
	(0.608)	(0.768)	(0.606)	(0.771)	(1.880)	(1.227)	(2.269)	(1.501)
Wales	1.803***	0.669	1.838***	0.689	-0.0938	3.434**	1.463	4.015**
	(0.655)	(0.853)	(0.653)	(0.858)	(1.994)	(1.623)	(2.416)	(1.888)
Scotland	1.144	0.487	1.226	0.471	-1.899	0.134	-1.505	0.281
	(0.769)	(0.825)	(0.761)	(0.825)	(1.803)	(1.791)	(2.126)	(1.966)
Northern Ireland	1.798	2.095*	1.722	2.158**	-2.845	0.476	-2.369	1.263
	(1.208)	(1.103)	(1.190)	(1.096)	(3.065)	(1.507)	(3.649)	(1.820)
Constant	23.58***	23.07***	23.40***	22.94***	26.14***	22.80***	24.87***	21.56***
	(0.500)	(0.676)	(0.498)	(0.680)	(1.705)	(1.052)	(2.046)	(1.316)
Observations	85,031	119,270	85,031	119,270	40,769	57,446	40,769	57,446
R-squared	0.024	0.031	0.025	0.032	0.112	0.140	0.023	0.030
AIC	455,517	675,926	455,372	675,732	206,677	308,464	210,580	315,350
BIC	455,919	676,342	455,839	676,217	207,073	308,876	210,942	315,726
Number of individuals	7,343	10,113	7,343	10,113	7,343	10,113	7,343	10,113

*Notes: Estimation 1 incorporates 10 pre-COVID-19 waves and no structural break. Estimation 2 incorporates 10 pre-COVID-19 waves. Estimation 3 incorporates 2 pre-COVID-19 waves and the variables associated with loneliness. Estimation 4 incorporates 2 pre-COVID-19 waves. Clustered-robust standard errors in parentheses; *p-value < 0.1, **p-value < 0.05, ***p-value < 0.01. R-squared refers to the within R-squared as reported by Stata. The structural break in the coefficients is generated by interacting the variables for which the note '(during COVID-19)' is reported with a binary variable which distinguishes the time after March 2020 from the previous period. For the variables capturing housing tenure and region no structural break is assumed to occur.*

Table 2: Aggregate effect of each variable during the pandemic era based on the estimation in Table 1.

	2		**3**		**4**	
Variable	**Males**	**Females**	**Males**	**Females**	**Males**	**Females**
Not living with a partner	-0.731***	-0.206	0.232	0.257*	-0.179	-0.099
Child under 15 in the household	-0.684***	-0.535***	-0.241	-0.038	-0.276*	-0.116
How often feels lonely:						
Some of the time	-	-	-2.666***	-2.927***	-	-
Often	-	-	-7.241***	-7.720***	-	-
Long-standing health condition	0.022	0.067	0.050	0.262**	-0.022	0.269**
Employed:						
Yes	0.341	-0.128	0.094	0.237	0.251	0.170
Retired	1.144***	0.685***	0.561**	0.252	0.731**	0.518*
Hours worked per week	0.009***	0.009***	0.011***	0.006**	0.013***	0.013***
Absolute income	0.087***	0.075***	0.112***	0.072***	0.107***	0.059**

*Notes: *p-value < 0.1, **p-value < 0.05, ***p-value < 0.01. The associated test is a test for the significance of the aggregate effect of each variable during the pandemic era. Test is carried out using a two-sided Wald test with a null hypothesis stating that the sum of the coefficients associated with a variable in the pre-COVID-19 period and the structural break during the pandemic is equal to 0.*

4.2.1 Labour market impact

A notable feature of the pandemic is that it affected the job market in ways that were previously unprecedented, particularly with the introduction of the furlough scheme and home working. The furlough scheme was introduced in March 2020 as part of the Coronavirus Job Retention Scheme. At the start of the pandemic HMRC covered 80% of the wages of furloughed employees. This was done in an attempt to help avoid redundancies and was initially due to end on the 30th May 2020. The scheme was extended four times, with the level of government support changing. By the end of the scheme on the 30th September 2021, it paid 60%, with employers covering 20% of wages[14]. In addition, during March 2020, the Government issued a statement encouraging people to start working from home wherever possible. By April 2020 almost half of the working population in the UK did some work at home[15].

In what follows we expand the discussion by examining the effect furloughing, working from home, and self-isolating had on mental health by including a series of employment interaction effects. Four different models are estimated and presented in *Table 3*. As before the results are disaggregated in males and females. The first specification presents the original job-related arguments (***Model 1***). The second specification differentiates between those who were furloughed and those who were not (***Model 2***). The third accounts for the frequency of working from home (***Model 3***), and the last differentiates between those who were self-isolating and those who were not (***Model 4***). As before, *Table 4* presents the total effect of each variable on mental health, i.e. by summing across the relevant interaction effects for each subgroup.

Tables 3 and *4* show that in line with the literature there are no adverse effects on mental health from being furloughed; the interaction effect representing furloughed individuals is not significant. Ferry *et al.* (2021), for example, do not find a significant association between being furloughed and the level of psychological distress in April 2020. Chandola *et al.* (2020) find that up to July 2020 the probability that an individual experienced a common mental health disorder was similar among furloughed employees and those whose jobs were not affected. *Table 4* even suggests an overall positive impact for males when the reference group are individuals without a job (***Model 2***). In a model which controls for loneliness, furloughing is associated with a positive effect on mental health (see, column 1 of *Table E.1* in Appendix E),

[14] https://commonslibrary.parliament.uk/examining-the-end-of-the-furlough-scheme/.

[15] https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/bulletins/coronavirusandhomeworkingintheuk/april2020.

providing tentative evidence that the results in *Table 3* might be capturing the impact of being isolated at home which accompanies furloughing.

Accounting for the frequency of working from home (***Model 3***) suggests that for both males and females always working from home has a detrimental effect on mental health compared to never working from home[16]. Males who are employed and never work from home are the only group whose mental health is significantly higher relative to those who do not have a job, while females who are employed and always work from home have significantly worse mental health at the 10% level than those without a job. These results speak both to the preferences of individuals with regard to working from home during the pandemic, suggesting the need to leave the house at least for some of the working day, and possibly to the different roles men and women have in the home. In a model which controls for loneliness, the frequency of working from home is not as important among employed individuals, but this seems to be driven mainly by the different sample of observations used in the estimation (see, Appendix E).

Lastly, we also differentiate between individuals who have a job and those who are employed but are self-isolating (***Model 4***). Here we find that self-isolating (which is linked to being in poor health) has a negative and significant effect on mental health for both males and females relative to those employed and not self-isolating, a finding which agrees with Ferry *et al*. (2021). The overall effect is also negative and significant relative to those who do not have a job. These findings hold even for the model which incorporates loneliness (see, Appendix E).

In summary, the pre-pandemic positive association between mental health and employment is no longer significant during the pandemic. Aside from the general uncertainty in the job market, this may reflect changes in the nature of employment during the pandemic. For many their mental health became similar to those without a job. There is a well-being advantage from never working from home, while self-isolation worsens mental health. In addition, the furlough scheme does not generate adverse consequences in mental health, compared to those unaffected by the scheme.

[16] The full benefit of attending the workplace all the time may be hindered by the associated increased risk of exposure to the virus.

Table 3: Within estimator for job-related variables for years 2009-2021.

Variable	Males	Females
Model 1		
Employed (Default: No)		
Yes	1.156***	0.756***
	(0.140)	(0.103)
Retired	1.733***	1.575***
	(0.137)	(0.116)
Yes (during COVID-19)	-0.815***	-0.884***
	(0.255)	(0.186)
Retired (during COVID-19)	-0.589**	-0.890***
	(0.260)	(0.240)
Model 2		
Employed (Default: No)		
Yes	1.153***	0.755***
	(0.140)	(0.103)
Retired	1.732***	1.574***
	(0.137)	(0.116)
Yes (during COVID-19)	-0.868***	-0.911***
	(0.263)	(0.191)
Retired (during COVID-19)	-0.587**	-0.888***
	(0.260)	(0.240)
Furloughed during COVID-19 (Default: No)		
Furloughed	0.247	0.142
	(0.162)	(0.150)
Missing furlough state	0.0379	0.0233
	(0.114)	(0.118)
Model 3		
Employed (Default: No)		
Yes	1.153***	0.758***
	(0.139)	(0.103)
Retired	1.729***	1.571***
	(0.136)	(0.116)
Yes (during COVID-19)	-0.716***	-0.834***
	(0.258)	(0.188)
Retired (during COVID-19)	-0.586**	-0.888***
	(0.260)	(0.240)
Worked from home during COVID-19 if employed (Default: Never worked from home)		
Always worked from home	-0.330***	-0.232**
	(0.113)	(0.105)
Often worked from home	-0.227*	-0.0495
	(0.131)	(0.118)
Sometimes worked from home	-0.0512	-0.0576
	(0.118)	(0.113)
Missing work from home status	1.550	-1.989
	(1.351)	(1.750)

Model 4		
Employed (Default: No)		
Yes	1.156***	0.757***
	(0.139)	(0.103)
Retired	1.735***	1.576***
	(0.136)	(0.116)
Yes (during COVID-19)	-0.676***	-0.703***
	(0.254)	(0.187)
Retired (during COVID-19)	-0.592**	-0.882***
	(0.260)	(0.240)
Self-isolated during COVID-19 if employed (Default: Did not self-isolate)		
Self-isolated	-1.741***	-1.779***
	(0.308)	(0.297)
Missing self-isolating status	-0.0973	-0.180**
	(0.0773)	(0.0737)

*Notes: Clustered-robust standard errors in parentheses; *p-value < 0.1, **p-value < 0.05, ***p-value < 0.01. The specified model for each estimation is not reported and is the same as in Table 1 for estimation 2 apart from the job-related variables. In order to avoid dropping further observations for the COVID-19 version of Understanding Society, an additional dummy variable for missing values is added in all estimations for all new job-related variables.*

Table 4: Aggregate effect of each variable during the pandemic era based on the estimation in Table 3.

Variable	Males	Females
Model 1 **(Default: Unemployed)**		
Yes	0.341	-0.128
Retired	1.144***	0.685***
Model 2 **(Default: Unemployed)**		
Employed and not furloughed	0.285	-0.156
Employed and furloughed	0.532**	-0.014
Employed and missing furlough status	0.323	-0.133
Retired	1.145***	0.686***
Model 3 **(Default: Unemployed)**		
Employed and always works from home	0.107	-0.308*
Employed and often works from home	0.210	-0.126
Employed and sometimes works from home	0.386	-0.134
Employed and never works from home	0.437*	-0.076
Employed and missing work from home status	1.987	-2.065
Retired	1.143***	0.683***
Model 4 **(Default: Unemployed)**		
Employed and not self-isolating	0.480**	0.054
Employed and self-isolating	-1.261***	-1.725***
Employed and missing self-isolating status	0.383	-0.126
Retired	1.143***	0.694***

*Notes: *p-value < 0.1, **p-value < 0.05, ***p-value < 0.01. The associated test is a test for the significance of the aggregate effect of each variable during the pandemic era. Test is carried out using a two-sided Wald test with a null hypothesis stating that the sum of the coefficients associated with a variable in the pre-COVID-19 period and during the pandemic is equal to 0.*

4.2.2 Income-related variables

Next, we expand the discussion to include components which capture social comparisons of income. These include reference income and the rank of income. In line with other studies (see, subsection 2.2), each reference group is constructed by partitioning the sample within each time period[17] according to three characteristics [18]: 1) Region (12 regions); 2) Age (grouped into categories of <21, 21-25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60, 61-65, 66-70, >70); and 3) Education (6 levels).

The results are presented in *Tables 5* and *6* for males and females. Five different specifications are estimated. The first specification controls for absolute income and is identical to the original estimation (***Model 1***). ***Model 2*** includes absolute and reference income. ***Model 3*** includes all three. ***Model 4*** includes absolute income and the rank of income. Finally, ***Model 5*** only includes the rank of income. As before, *Table 6* presents the overall effect each variable has on mental health during the pandemic, calculated by summing across the relevant interaction effects. The five specifications chosen are based on the relevant literature (subsection 2.2) and taken together allow us to study the impact of the different social comparison mechanisms both collectively, and in isolation.

Table 5 shows that regardless of the specification, the income-related components have a significant effect on mental health in the pre-COVID-19 period for males and females. In line with other studies, higher absolute income and being of higher rank within the reference group both lead to better mental health (see, Clark *et al*., 2009; Boyce *et al*., 2010), whereas a higher reference income (implying a higher social comparison reference point) has a detrimental effect on mental health (see, Ferrer-i-Carbonell, 2005; Becchetti *et al*., 2013). Such a finding reflects the co-importance of the two social comparison mechanisms for mental health before the pandemic. In the pre-COVID-19, people care both about the general income level of their peers, as well as how they rank among them.

For males, there is no evidence that the positive impact of absolute income changes during the pandemic (*Table 5*); the overall significant positive effect continues to hold across all

[17] The time period considered refers to the calendar year for each of the ten waves coming from the original survey of Understanding Society, and the calendar month for each of the nine waves coming from the COVID-19 version. It should be noted that the first three months of 2020 which are recorded in wave 10 of the original survey are included in the reference group for 2019 in order to keep each time period fairly balanced in terms of the sample size incorporated in the construction of reference groups.

[18] The reference group constructed for each individual uses observations for which income and the relevant demographic variables are not missing. This results in using more than the 204,301 observations used in the analysis and is done such that the full availability of the sample distributions is used.

specifications (*Table 6*). Similarly, reference income has a significant negative effect during the pandemic, which is reinforced relative to the pre-pandemic period (*Table 5*). On the other hand, the overall impact of the rank of income appears to diminish, as conveyed in *Table 6*. Rank only has an overall significant effect during the pandemic when included as the sole income-related component. However, in this case it is also capturing the impact of absolute income itself as the two variables are highly correlated.

We also attempt to use a general-to-specific modelling approach for the income-related variables. The general-to-specific approach involves starting with a general model which involves all variables which are potentially important, and through a stepwise procedure removes the ones which seem unimportant empirically. We therefore start from ***Model 3*** which includes all income-related components with structural breaks. We end up with a model which excludes the impact of rank during the pandemic and includes structural breaks for both absolute income and reference income. This estimated model has a lower AIC (455,349) than any of the models presented in this subsection, and a BIC (455,844) which is close to the minimum[19]. This result suggests that the only social comparison mechanism which affects the mental health of males during the pandemic is reference income. Its impact, along with that of absolute income, are greater in magnitude after the onset of COVID-19 in the aforementioned estimation. Results available on request.

For females, the overall significance of absolute income during the pandemic varies depending on whether the rank of income is included in the model (*Table 6*). In ***Models 3*** and ***4*** in which the rank of income is included along with the absolute income, the rank of income is the only significant income-related component during the pandemic. The overall effect of reference income becomes insignificant during the pandemic period (*Table 6*). In contrast, the overall significant impact of rank of income persists during the pandemic (*Table 6*), albeit not significantly changing (*Table 5*).

Once again, by using a general-to-specific modelling approach starting from ***Model 3***, which includes all income-related components with structural breaks, a model which removes the impact of reference income during the pandemic, and incorporates absolute income and rank of income without a structural break has lower AIC (675,700) and BIC (676,194) values than any of the models presented in *Table 5*. This result suggests that the rank of income is the only

[19] Income-related components are correlated with each other, and this can hinder the inferences made when they are all included in the same model. As such, non-nested model comparison statistics, such as AIC and BIC, might be useful in comparing competing social comparison mechanisms when used in isolation.

social comparison mechanism which matters for the mental health of females during the pandemic. Results available on request.

Overall, the results above convey that males seem to move away from the rank social comparison mechanism during the pandemic, while females move away from the reference income mechanism. In both cases, however, social comparison concerns persist. Males appear to care more about the general income level of their peers, and not necessarily how they rank in their reference group, whereas females care more about where they rank among their peers, regardless of their level of income. For both males and females, the absolute income which reflects the standard of living, is still significant for mental health during the pandemic. Therefore, any differences in social comparisons might reflect inherent characteristics of males and females on how they determine their relative position in their reference group during periods of crisis.

Table 5: Within estimator for income-related variables for years 2009-2021.

Variable	Males	Females
Model 1		
Absolute income	0.0520***	0.0555***
	(0.0141)	(0.0132)
Absolute income (during COVID-19)	0.0349	0.0193
	(0.0292)	(0.0279)
AIC	455,372	675,732
BIC	455,839	676,217
Model 2		
Absolute income	0.0550***	0.0603***
	(0.0142)	(0.0133)
Absolute income (during COVID-19)	0.0452	0.0126
	(0.0299)	(0.0283)
Reference income	-0.0674***	-0.0979***
	(0.0251)	(0.0287)
Reference income (during COVID-19)	-0.292**	0.0956
	(0.132)	(0.127)
AIC	455,355	675,725
BIC	455,842	676,228
Model 3		
Absolute income	0.0383**	0.0410***
	(0.0157)	(0.0149)
Absolute income (during COVID-19)	0.0645*	-0.00638
	(0.0351)	(0.0334)
Reference income	-0.0492*	-0.0734**
	(0.0256)	(0.0294)
Reference income (during COVID-19)	-0.322**	0.117
	(0.135)	(0.130)
Rank of income	0.232**	0.298***
	(0.0974)	(0.0973)
Rank of income (during COVID-19)	-0.223	0.0456
	(0.185)	(0.166)
AIC	455,351	675,706
BIC	455,856	676,229
Model 4		
Absolute income	0.0355**	0.0342**
	(0.0154)	(0.0145)
Absolute income (during COVID-19)	0.0421	0.00587
	(0.0338)	(0.0324)
Rank of income	0.251***	0.343***
	(0.0949)	(0.0949)
Rank of income (during COVID-19)	-0.141	-0.0134
	(0.181)	(0.162)
AIC	455,365	675,709
BIC	455,851	676,212

Model 5		
Rank of income	0.356***	0.440***
	(0.0870)	(0.0859)
Rank of income (during COVID-19)	-0.0281	-0.00412
	(0.158)	(0.141)
AIC	455,382	675,717
BIC	455,849	676,201

*Notes: Clustered-robust standard errors in parentheses; *p-value < 0.1, **p-value < 0.05, ***p-value < 0.01. The specified model for each estimation is not reported and is the same as in Table 1 for estimation 2 apart from the income-related variables. Rank of income refers to the Decision-by-Sampling component mentioned in subsection 2.2. The structural break in the coefficients is generated by interacting the variables for which the note '(during COVID-19)' is reported with a binary variable which distinguishes the time after March 2020 from the previous period.*

Table 6: Aggregate effect of each variable during the pandemic era based on the estimation in Table 5.

Variable	Males	Females
Model 1		
Absolute income	0.0869***	0.0748***
Model 2		
Absolute income	0.100***	0.0729***
Reference income	-0.359***	-0.0023
Model 3		
Absolute income	0.103***	0.0346
Reference income	-0.371***	0.0436
Rank of income	0.009	0.344**
Model 4		
Absolute income	0.0776**	0.0401
Rank of income	0.110	0.330**
Model 5		
Rank of income	0.328**	0.436***

*Notes: *p-value < 0.1, **p-value < 0.05, ***p-value < 0.01. The associated test is a test for the significance of the aggregate effect of each variable during the pandemic era. Test is carried out using a two-sided Wald test with a null hypothesis stating that the sum of the coefficients associated with a variable in the pre-COVID-19 period and during the pandemic is equal to 0.*

4.3 Robustness checks

We turn now to examine the timing of the structural break. In doing so, we examine how sensitive our results are to alternative assumptions about when the structural break occurred, plus whether there are any subsequent breaks.

4.3.1 Timing of the first structural break

The models so far have been estimated under the assumption that the structural break coincides with the implementation of the first lockdown in the UK. However, it may very well be the case that the structural break did not coincide precisely with this date. We investigate this in what follows by varying the date of the structural break. Each month during which data is recorded for Understanding Society's COVID-19 web survey is used as a candidate structural break[20]. We then compare the original AIC/BIC statistics with the new specifications.

As a rule of thumb, Fabozzi *et al.* (2014) suggest that a difference in the AIC/BIC statistics of 2 or more provides evidence against the model with the higher AIC/BIC statistic[21]. As already mentioned, lower AIC and BIC values act as indication of a better fit with the data. The results are presented in *Table 7* for males and females, respectively. In column 1 we estimate the original model, while in column 2 we also control for loneliness. By definition, the AIC difference is identical to the BIC difference between any two models with the same number of parameters, and thus only one value is reported per period.

The results show that the estimated models which use the break dates after March 2020 have a difference in AIC/BIC from the original model which is substantially larger than 2, providing evidence against the alternative break dates.

[20] This implies that the timing of any structural break cannot be pinpointed exactly with respect to the month, but is subject to data availability. Despite this, the detection of a structural break is still possible.
[21] The authors also mention more sophisticated measures of model comparison which are based on the relative AIC and BIC differences between models, namely *Akaike weights* and *evidence ratios*. However, in this case they add no value to the analysis as the eventual model selection remains unchanged.

Table 7: Differences in AIC/BIC statistics for models with different structural break dates relative to original model.

	1		2	
Structural break date	**Males**	**Females**	**Males**	**Females**
March 2020	0	0	0	0
April 2020	42	57	121	164
May 2020	42	77	164	183
June 2020	70	94	208	188
July 2020	85	98	244	213
September 2020	83	110	214	245
November 2020	97	128	237	300
January 2021	121	165	261	323
March 2021	118	178	257	338

Notes: The structural break date mentioned in the table refers to the last month assumed to be part of the regime associated with the pre-COVID-19 period. By definition, the AIC difference is identical to the BIC difference between any two models with the same number of parameters, and thus only one value is reported per period. Estimation 1 incorporates 10 pre-COVID-19 waves. Estimation 2 incorporates 2 pre-COVID-19 waves and the variables associated with loneliness.

4.3.2 Second structural break

In an uncertain period like the pandemic, it may be the case that a second structural break also occurred. To examine this we apply the following procedure. First, we assume the first structural break date to be March 2020, and then estimate a set of models each time varying the date of a second structural break. We examine dates from April 2020 to March 2021. This is done for the model which incorporates feelings of loneliness as well. The models estimated have the form:

$$GHQ_{it} = \alpha + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \boldsymbol{z}'_{it}\boldsymbol{\gamma_1} + 1(c < t \leq k)\boldsymbol{z}'_{it}\boldsymbol{\gamma_2} + 1(t > k)\boldsymbol{z}'_{it}\boldsymbol{\gamma_3} + d_t + h_i + \varepsilon_{it}, \qquad (3)$$

where every term is defined exactly as in equation (2), apart from parameter k which represents the timing of the second structural break. As such, the impact of $\boldsymbol{z}_{it}$ on mental health is captured by the vector of coefficients $\boldsymbol{\gamma_1} + \boldsymbol{\gamma_2}$ in the period between the first and second breaks, and by the vector $\boldsymbol{\gamma_1} + \boldsymbol{\gamma_3}$ after the second structural break.

The models with the lowest AIC/BIC statistics are chosen and for each one the significance of the second structural break is examined. This is done by using a Wald test with a joint null hypothesis that the effect of each variable is the same in the period between the first and second breaks in relation to the period after the second break. This is equivalent to testing that $\boldsymbol{\gamma_2} = \boldsymbol{\gamma_3}$ in specification (3). The AIC/BIC statistics of each model are presented in *Table 8*, for males and females, respectively.

Column 1 shows that, for the model without loneliness, there is not clear evidence of a second structural break. The AIC statistic favours the model with a second structural break, September 2020 for males and July 2020 for females, whereas the BIC statistic favours the model without a second structural break. In addition, the p-values for the significance test of the second structural break are marginally significant at 0.061 and 0.014 for males and females, respectively.

In the case of incorporating the loneliness aspect, the tests are clearly in favour of a second structural break with p-values at 0.000 for both males and females. June or July 2020 is the selected break date for males, and June 2020 for females. For males, the BIC statistic still favours the model without a second structural break, but the gap in BIC statistics between the one-break and two-break models is substantially reduced relative to the case of not incorporating the loneliness variable.

Table 8: AIC/BIC statistics for models with different dates for second structural break.

	1		2	
Second structural break date	**Males**	**Females**	**Males**	**Females**
No second structural break				
AIC	455,372	675,732	206,677	308,464
BIC	455,839	676,217	207,073	308,876
April 2020				
AIC	455,375	675,735	206,673	308,430
BIC	455,908	676,287	207,147	308,923
May 2020				
AIC	455,369	675,730	206,646	308,383
BIC	455,902	676,282	207,120	308,876
June 2020				
AIC	455,371	675,723	206,625	208,346
BIC	455,904	676,275	207,099	308,839
July 2020				
AIC	455,366	675,718	206,625	308,386
BIC	455,899	676,270	207,099	308,879
September 2020				
AIC	455,359	675,724	206,640	308,427
BIC	455,892	676,277	207,113	308,920
November 2020				
AIC	455,361	675,730	206,658	308,456
BIC	455,894	676,282	207,132	308,949
January 2021				
AIC	455,374	675,732	206,679	308,459
BIC	455,907	676,284	207,153	308,952
March 2021				
AIC	455,362	675,732	206,671	308,461
BIC	455,895	676,284	207,145	308,954

Notes: The structural break date mentioned in the table refers to the last month assumed to be part of the regime associated with period between the first and second breaks. Estimation 1 incorporates 10 pre-COVID-19 waves. Estimation 2 incorporates 2 pre-COVID-19 waves and the variables associated with loneliness.

Using the dates with the lowest AIC/BIC statistics for the second structural break, models with two structural breaks each are estimated. The results are in Appendix F. As before the models are estimated with and without loneliness. The results show that the second structural break mainly reflects a reduction in the mental health burden of those reporting feelings of loneliness 'Some of the time'. For females, absolute income is not significant in the period between the first and second structural breaks. However, after the second structural break the absolute income variable reverts back to a significant positive impact on mental health. The rest of the variables do not exhibit any significant changes across the pandemic period.

Finally, as an additional robustness check for the second structural break, we also follow the approach outlined in Hansen (1999). This approach treats the timing of the structural break as

a nuisance parameter to be estimated. Hansen (1999) proposes a test with a null hypothesis of no structural break to check the significance of the estimated break[22]. The approach extends to the estimation and testing of more than one structural break[23]. The approach suggested by Hansen (1999) can be applied only to balanced panel data sets. Since Understanding Society is an unbalanced panel, if we used the longest time span available (2009 to 2021) we would only have 19,361 observations (1,019 individuals). Instead, we reduce the time span by restricting the pre-COVID-19 period to the last 4 years of data. This gives us a balanced sample of 30,966 observations (2,382 individuals). This approach is also applied to the data on loneliness, which gives us a balanced sample comprising of 29,469 observations from 2,679 individuals. In order to avoid reducing the sample size even further we do not split the sample between males and females. The reduced sample size is comparable to that used by Wang (2015) who carries out a simulation-based evaluation of the performance of Hansen's approach (1999)[24].

In the model without loneliness no significant second structural break is found (the relevant p-value is 0.640)[25]. For the model with loneliness a significant second structural break is found (relevant p-value is 0.000) in June 2020[26]. This is in agreement with the structural break found through the previous approach[27]. These results are in Appendix G.

[22] The critical values against which the test statistic is compared are constructed through a bootstrap design proposed by Hansen (1996).

[23] In the case of investigating multiple structural breaks, a sequential estimator is applied which is found to be consistent based on the works of Bai (1997), and Bai and Perron (1998). In particular, a second structural break date is estimated given the estimate of the first structural break date. The same goes for the estimation of a third structural break date given the estimates of the other two. The testing procedure is also sequential. As such, the testing should proceed until the last structural break investigated is not significant.

[24] Hansen's (1999) method is implemented by using the Stata command `xthreg` developed by Wang (2015). The possibility of the detection of up to three structural breaks is offered in the Stata package. Structural breaks are investigated in terms of the coefficients of the variables for which the structural break is assumed in *Table 1*. The critical values for the structural break significance tests are generated through bootstrapping with 300 replications. Clustered-robust standard errors are assumed for the estimated models.

[25] This is true even in the case of splitting the sample in males and females.

[26] An insignificant third structural break is found in September 2020 (relevant p-value is 0.633).

[27] When Hansen's (1999) approach is applied to males and females separately, June 2020 is detected as a structural break in both cases (relevant p-values are 0.373 for males and 0.030 for females). Given the significance of June 2020 detected for males by using the first approach, it is not clear if the insignificance in this case is due to the severe reduction in sample size or due to the fact that there is indeed no second structural break for males under Hansen's (1999) approach.

5. CONCLUSION

The onset of the COVID-19 pandemic in the UK was daunting for many aspects of life, including the significant deterioration in mental health of the general population. Many of the key determinants of mental health were significantly changed during the pandemic relative to the pre-pandemic period. Therefore, a portion of the reduction in the general level of mental health was predictable and has been reported in the literature. However, it is also important to examine whether there has been any structural change with respect to how each aspect of life influences the level of mental health. This can provide evidence on how mental health is determined during periods of crisis which can then be used by policy makers to ensure safety nets are appropriately targeted. A structural break is detected in mental health determination with respect to core socioeconomic aspects in the lives of individuals.

The aforementioned socioeconomic aspects include cohabitation, whether or not there is a school-aged child in the household, employment, loneliness feelings, health, hours worked per week, and absolute income. Apart from the absolute level of income, social comparison components are also incorporated in the analysis. The reference income is included which represents the average income of a group of individuals who are similar according to certain socio-demographic characteristics. The rank of income is also included which is the normalized ranking of an individual's income within the aforementioned group of similar individuals.

A model incorporating a single structural break is estimated for both males and females. The date of the structural break is assumed to be known as March 2020 during which the first national lockdown was imposed in the UK. In the context of a single structural break, living with a partner is originally associated with a positive structural break effect on mental health during the pandemic. However, this is capturing the negative association between the level of mental health and the frequency of feeling lonely which appears to be reinforced after the onset of COVID-19 for both males and females. There is also evidence of increased mental distress associated with having a child aged 15 or under in the household. Lastly, the case of having a long-term health condition which is associated with a higher level of mental distress in the pre-pandemic period is found to be linked with a structural change in the opposite direction during the pandemic.

For job-related variables, the mental health premium associated with being employed or retired as opposed to being unemployed in the pre-pandemic period is found to be significantly reduced during the pandemic. Furthermore, the number of hours worked per week obtains a

positive association with the level of mental health during the pandemic as opposed to the negative one before. In examining different features of the labour market, a well-being premium associated with never working from home is found, while self-isolation reflects lower mental health among those employed. Lastly, the furlough scheme does not involve adverse effects on mental health, compared to those unaffected by the scheme.

In the context of the single known structural break date, the specification with respect to income-related arguments is also investigated. Social comparison concerns in the pre-pandemic period persist after the onset of COVID-19, albeit through different social comparison mechanisms for males and females. For males, the average income of others matters. For females, their income relative to others in the form of a ranking matters.

We also find tentative evidence of a second structural break during the summer of 2020, after many of the UK's COVID-19 restrictions had been eased. This is mainly caused by the reduced mental health burden of those experiencing heightened feelings of loneliness.

Overall, there is evidence to support the existence of at least one structural break in the mental health determination equation. This is a structural break in the association between vital aspects of every-day life and the mental health of individuals. An interesting topic for further research is the investigation of whether the aforementioned relationships shift back to their pre-pandemic state as more data becomes available with time, or if there is some form of permanent structural change as a result of how the unprecedented circumstances have influenced the perception of individuals in modern UK. Also, it is interesting to consider whether these changes are also evident in other countries of the world. It is vital to understand how COVID-19 impacted mental health determination, and whether focus should be placed on aspects other than those considered before the pandemic, at least for the short run. Evidence in the current study suggest that policy makers should aim at policies against job uncertainty as the mental health gap between employed and unemployed individuals was significantly reduced after the structural break, and working hours now exhibit a positive significant association with the level of mental health. Policies implemented by the UK government such as financial support, or the furlough scheme are supported by the current analysis in the context of recovering from mental health deterioration. It remains to be seen whether or how long it will take for the average level of metal health to fully recover, and what the mental health determination equation will look like at that point.

6. REFERENCES

Angrist, J. and Pischke, J., 2009. *Mostly harmless econometrics: an empiricist's companion*. Princeton: Princeton University Press.

Baek, I. and Jun, J., 2011. Testing contagion of the 1997–98 crisis in Asian stock markets with structural breaks and incubation periods. *Journal of Asian Economics*, 22(5), pp.356-368.

Bai, J. and Perron, P., 1998. Estimating and Testing Linear Models with Multiple Structural Changes. Econometrica, 66(1), pp.47-78.

Bai, J., 1997. Estimating Multiple Breaks One at a Time. Econometric Theory, 13(3), pp.315-352.

Banks, J. and Xu, X., 2020. The Mental Health Effects of the First Two Months of Lockdown during the COVID-19 Pandemic in the UK*. *Fiscal Studies*, 41(3), pp.685-708.

Banks, J., Fancourt, D. and Xu, X., 2021. In: H. John, R. Layard, J. Sachs and J. De Neve, *World Happiness Report 2021*. New York: Sustainable Development Solutions Network, pp.107-130.

Becchetti, L., Castriota, S., Corrado, L. and Ricca, E., 2013. Beyond the Joneses: Inter-country income comparisons and happiness. *The Journal of Socio-Economics*, 45, pp.187-195.

Boes, S., Staub, K. and Winkelmann, R., 2010. Relative status and satisfaction. *Economics Letters*, 109(3), pp.168-170.

Boyce, C., Brown, G. and Moore, S., 2010. Money and Happiness. *Psychological Science*, 21(4), pp.471-475.

Brown, S., Gray, D. and Roberts, J., 2015. The relative income hypothesis: A comparison of methods. *Economics Letters*, 130, pp.47-50.

Bu, F., Steptoe, A. and Fancourt, D., 2020. Who is lonely in lockdown? Cross-cohort analyses of predictors of loneliness before and during the COVID-19 pandemic. *Public Health*, 186, pp.31-34.

Cameron, A. and Trivedi, P., 2005. *Microeconometrics*. Cambridge: Cambridge University Press.

Card, D., Mas, A., Moretti, E. and Saez, E., 2012. Inequality at Work: The Effect of Peer Salaries on Job Satisfaction. *American Economic Review*, 102(6), pp.2981-3003.

Chandola, T., Kumari, M., Booker, C. and Benzeval, M., 2020. The mental health impact of COVID-19 and lockdown-related stressors among adults in the UK. *Psychological Medicine*, pp.1-10.

Clark, A. and Oswald, A., 1996. Satisfaction and comparison income. *Journal of Public Economics*, 61(3), pp.359-381.

Clark, A. and Oswald, A., 2002. A simple statistical method for measuring how life events affect happiness. *International Journal of Epidemiology*, 31(6), pp.1139-1144.

Clark, A., 2018. Four Decades of the Economics of Happiness: Where Next?. *Review of Income and Wealth*, 64(2), pp.245-269.

Clark, A., Kristensen, N. and Westergård-Nielsen, N., 2009. Economic Satisfaction and Income Rank in Small Neighbourhoods. Journal of the European Economic Association, 7(2-3), pp.519-527.

Crossley, T., Fisher, P. and Low, H., 2021. The heterogeneous and regressive consequences of COVID-19: Evidence from high quality panel data. *Journal of Public Economics*, 193, pp.104334-104334.

Daly, M. and Robinson, E., 2021. Longitudinal changes in psychological distress in the UK from 2019 to September 2020 during the COVID-19 pandemic: Evidence from a large nationally representative study. *Psychiatry Research*, 300, pp.113920-113920.

Daly, M., Boyce, C. and Wood, A., 2015. A social rank explanation of how money influences health. *Health Psychology*, 34(3), pp.222-230.

Daly, M., Sutin, A. and Robinson, E., 2020. Longitudinal changes in mental health and the COVID-19 pandemic: evidence from the UK Household Longitudinal Study. *Psychological Medicine*, pp.1-10.

Duesenberry, J., 1949. *Income, saving, and the theory of consumer behavior*. Cambridge (Massachusetts): Harvard University Press.

Easterlin, R. 1974. Does economic growth improve the human lot? Some empirical evidence. In: P. David and M. Reder, *Nations and Households in Economic Growth: Essays in Honor of Moses Abramovitz*. New York and London: Academic Press, pp.89-125.

Ellwardt, L. and Präg, P., 2021. Heterogeneous mental health development during the COVID-19 pandemic in the United Kingdom. *Scientific Reports*, 11.

Fabozzi, F., Focardi, S., Rachev, S., Arshanapalli, B. and Höchstötter, M., 2014. *The Basics of Financial Econometrics*. Hoboken (New Jersey): John Wiley & Sons.

Fan, Y. and Xu, J., 2011. What has driven oil prices since 2000? A structural change perspective. *Energy Economics*, 33(6), pp.1082-1094.

Ferrer-i-Carbonell, A. and Frijters, P., 2004. How Important is Methodology for the Estimates of the Determinants of Happiness?. *The Economic Journal*, 114(497), pp.641-659.

Ferrer-i-Carbonell, A., 2005. Income and well-being: an empirical analysis of the comparison income effect. *Journal of Public Economics*, 89(5), pp.997-1019.

Ferrer-i-Carbonell, A., 2013. Happiness economics. *SERIEs*, 4(1), pp.35-60.

Ferry, F., Bunting, B., Rosato, M., Curran, E. and Leavey, G., 2021. The impact of reduced working on mental health in the early months of the COVID-19 pandemic: Results from the Understanding Society COVID-19 study. *Journal of Affective Disorders*, 287, pp.308-315.

Gerlach, R., Wilson, P. and Zurbruegg, R., 2006. Structural breaks and diversification: The impact of the 1997 Asian Financial Crisis on the integration of Asia-Pacific Real Estate Markets. *Journal of International Money and Finance*, 25(6), pp. 974-991.

Giovanis, E. and Ozdamar, O., 2021. Implications of COVID-19: The Effect of Working From Home on Financial and Mental Well-Being in the UK. *International Journal of Health Policy and Management*.

Goldberg, D., Gater, R., Sartorius, N., Ustun, T., Piccinelli, M., Gureje, O. and Rutter, C., 1997. The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychological Medicine*, 27(1), pp.191-197.

Hansen, B., 1996. Inference When a Nuisance Parameter Is Not Identified Under the Null Hypothesis. Econometrica, 64(2), pp.413-430.

Hansen, B., 1999. Threshold effects in non-dynamic panels: Estimation, testing, and inference. Journal of Econometrics, 93(2), pp.345-368.

Li, L. and Wang, S., 2020. Prevalence and predictors of general psychiatric disorders and loneliness during COVID-19 in the United Kingdom. *Psychiatry Research*, 291, pp.113267-113267.

Liu, M., Dufour, G., Sun, Z., Galante, J., Xing, C., Zhan, J. and Wu, L., 2021. The impact of the COVID-19 pandemic on the mental health of young people: A comparison between China and the United Kingdom. *Chinese Journal of Traumatology*, 24(4), pp.231-236.

Martins, A., Serra, A., Martins, F. and Stevenson, S., 2021. EU housing markets before financial crisis of 2008: The role of institutional factors and structural breaks. *Journal of Housing and the Built Environment*, 36(3), pp.867-899.

Niedzwiedz, C., Benzeval, M., Hainey, K., Leyland, A. and Katikireddi, S., 2021a. Psychological distress among people with probable COVID-19 infection: analysis of the UK Household Longitudinal Study. *BJPsych Open*, 7(3), pp.e104, 1-3.

Niedzwiedz, C., Green, M., Benzeval, M., Campbell, D., Craig, P., Demou, E., Leyland, A., Pearce, A., Thomson, R., Whitley, E. and Katikireddi, S., 2021b. Mental health and health behaviours before and during the initial phase of the COVID-19 lockdown: longitudinal analyses of the UK Household Longitudinal Study. *Journal of Epidemiology and Community Health*, 75(3), pp.224-231.

Oecd-ilibrary.org. 2011. *Divided We Stand*. [online] Available at: https://www.oecd-ilibrary.org/social-issues-migration-health/the-causes-of-growing-inequalities-in-oecd-countries_9789264119536-en.

Osafo Hounkpatin, H., Wood, A., Brown, G. and Dunn, G., 2014. Why does Income Relate to Depressive Symptoms? Testing the Income Rank Hypothesis Longitudinally. *Social Indicators Research*, 124(2), pp.637-655.

Pfaff, T., 2013. Income Comparisons, Income Adaptation, and Life Satisfaction: How Robust are Estimates from Survey Data?. *SSRN Electronic Journal*.

Pierce, M., Hope, H., Ford, T., Hatch, S., Hotopf, M., John, A., Kontopantelis, E., Webb, R., Wessely, S., McManus, S. and Abel, K., 2020. Mental health before and during the COVID-19 pandemic: a longitudinal probability sample survey of the UK population. *The Lancet Psychiatry*, 7(10), pp.883-892.

Pierce, M., McManus, S., Hope, H., Hotopf, M., Ford, T., Hatch, S., John, A., Kontopantelis, E., Webb, R., Wessely, S. and Abel, K., 2021. Mental health responses to the COVID-19 pandemic: a latent class trajectory analysis using longitudinal UK data. *The Lancet Psychiatry*, 8(7), pp.610-619.

Quintana-Domeque, C. and Proto, E., 2022. On the Persistence of Mental Health Deterioration during the COVID-19 Pandemic by Sex and Ethnicity in the UK: Evidence

from Understanding Society. *The B.E. Journal of Economic Analysis & Policy*, 22(2), pp.361-372.

Ravallion, M., 2017. A concave log-like transformation allowing non-positive values. *Economics Letters*, 161, pp.130-132.

Senik, C., 2004. When information dominates comparison: Learning from Russian subjective panel data. *Journal of Public Economics*, 88(9), pp.2099-2123.

Stewart, N., Chater, N. and Brown, G., 2006. Decision by sampling. *Cognitive Psychology*, 53(1), pp.1-26.

Wang, Q., 2015. Fixed-Effect Panel Threshold Model using Stata. The Stata Journal: Promoting communications on statistics and Stata, 15(1), pp.121-134.

Wood, A., Boyce, C., Moore, S. and Brown, G., 2012. An evolutionary based social rank explanation of why low income predicts mental distress: A 17 year cohort study of 30,000 people. *Journal of Affective Disorders*, 136(3), pp.882-888.

APPENDIX A

Table A.1: Variable definitions in harmonized data set.

Variable	Definition
GHQ	General Health Questionnaire. The range is from 0 to 36. A high value represents a high level of mental health.
Not living with a partner	Dummy variable taking the value of 1 if not living with a partner (husband/wife/civil partner/partner/cohabitee), 0 otherwise.
Child under 15 in the household	Dummy variable taking the value of 1 if there is a at least 1 child aged 15 or under in the household.
How often feels lonely	How often the individual feels lonely. Categorical variable including the cases of hardly ever or never, some of the time, and often.
Employed	Categorical variable including the cases of employed or self-employed, retired, and none of the two.
Furloughed	Dummy variable taking the value of 1 if employed individual participates in furlough scheme.
Worked from home	Categorical variable including the frequencies of never, sometimes, often, and always.
Self-isolated	Dummy variable taking the value of 1 if employed individual is self-isolating.
Housing tenure	Categorical variable including the cases of owned, owned with mortgage, rented, and other.
Government office region	Categorical variable including the cases of North East, North West, Yorkshire and the Humber, East Midlands, West Midlands, East of England, London, South East, South West, Wales, Scotland, and Northern Ireland.
Long-standing health condition	Dummy variable taking the value of 1 if the individual has long-standing health condition (asthma, arthritis, congestive heart failure, coronary heart disease, angina, heart attack or myocardial infarction, stroke, emphysema, hypothyroidism or an under-active thyroid, chronic bronchitis, any kind of liver condition, cancer or malignancy, diabetes, epilepsy, high blood pressure/hypertension, multiple sclerosis, other long-standing/chronic condition), 0 otherwise.
Hours worked per week	Number of working hours per week.
Absolute income	Inverse hyperbolic sine transformation of equivalised monthly net household labour income adjusted for inflation.
Reference income	Average absolute income of the reference group.
Rank of income	Normalized income ranking of each individual within the reference group.

Notes: For the COVID-19 version of Understanding Society the loneliness question has an additional specification in that it asks the respondents about their loneliness feelings in the last 4 weeks preceding the survey. Housing tenure is only recorded for waves 4, 6, 8, and 9 of the COVID-19 version of the data set. For the rest of the available waves, housing tenure for each individual is assumed to be the latest available observation. For wave 1 of the COVID-19 version the relevant variables for the construction of the health variable are not observed. As such, the so-called baseline health conditions are used for wave 1. Baseline health conditions refer to the conditions mentioned by individuals at the start of the COVID-19 survey for any period before that. Given the nature of the conditions in Table A.1, baseline can be a trustworthy replacement for contemporaneous. The monthly net household labour income is adjusted for inflation based on UK inflation data available by the Office for National Statistics, and equivalised as proposed by Pfaff (2013) in order to account for household size. The square root scale is used in that the household income is divided by the square root of the household size (see, OECD, 2011). It should be noted that the household size is not observed directly, and cannot be calculated exactly, for the first wave of the COVID-19 study. As such, the household size from the second wave which is one month apart is used. The equivalised net household labour income sample distribution resembles a highly non-normal distribution. A reasonable approach towards making it look more normally distributed is using the natural logarithm of income. However, the log-like transformation known as the inverse hyperbolic sine transformation is used instead as it allows preserving those observations for which individuals report a household labour income of 0 (see, Ravallion, 2017).

APPENDIX B

Table B.1: Sample within standard deviation of the variables considered in the examination of a structural break in the well-being determination pre- and during COVID-19.

	All		Males		Females	
Variable	**Before**	**During**	**Before**	**During**	**Before**	**During**
GHQ	3.63	3.37	3.32	2.97	3.84	3.62
Not living with a partner (=1)	0.17	0.12	0.16	0.12	0.18	0.12
Child under 15 in the household (=1)	0.21	0.12	0.21	0.12	0.22	0.12
How often feels lonely:						
Hardly ever or never	0.22	0.29	0.20	0.26	0.23	0.30
Some of the time	0.24	0.31	0.22	0.28	0.25	0.33
Often	0.13	0.16	0.11	0.13	0.14	0.17
Has a long-standing health condition (=1)	0.23	0.12	0.23	0.11	0.23	0.12
Employed:						
Yes	0.25	0.12	0.24	0.12	0.26	0.13
No	0.21	0.12	0.19	0.12	0.22	0.12
Retired	0.18	0.03	0.19	0.03	0.17	0.02
Hours worked per week	10.75	9.50	11.33	10.24	10.32	8.94
Absolute income	1.58	0.78	1.60	0.78	1.57	0.78
Observations	138,851	65,450	58,130	26,901	80,721	38,549

Notes: The statistics are calculated for a sample of 204,301 observations from 17,456 individuals. Individuals participating only once in the pre-COVID-19 survey or once in the COVID-19 survey can still provide valid observations in terms of the criteria for the sample used in model estimation in the results section. Therefore, such observations will offer no variation for the statistics generated in Table B.1. For a variable X with realisation x_{it} for individual $i \in N = \{1, \dots, n\}$ at time $t \in T_i = \{1, .., t_i\}$ the within sample variance s_w^2 based on which Table B.1 is constructed is given by the following formula $s_w^2 = \frac{1}{\sum_{i \in N} \sum_{t \in T_i} 1 - 1} \sum_{i \in N} \sum_{t \in T_i} (x_{it} - \bar{x}_i)^2$, where $\bar{x}_i = \frac{1}{|T_i|} \sum_{t \in T_i} x_{it}$. The loneliness variable is recorded only for the last two waves of the original survey before COVID-19. As such, there are only 32,765 observations for the variable associated with loneliness before COVID-19 (13,868 for males; 18,897 for females).

Table B.1 presents the sample within standard deviation both before (i.e. prior to March 2020) and during the COVID-19 pandemic for the main variables of interest. This is done in order to offer a reference point against which the variation for the pandemic period can be compared with, given that the pre-COVID-19 period spans almost 12 calendar years whereas the COVID-19 period spans only 18 months. For the categorical variables, a set of dummy variables is constructed and the sample within standard deviation for each dummy is reported.

For both males and females, the within standard deviation for feelings of mental distress during the pandemic appears to be comparable in size with the pre-COVID-19 one.

The variation in feelings of loneliness is greater during the pandemic for both males and females. Given the turbulence and uncertainty of the pandemic, from both a social and economic perspective, this change can be intuitively expected.

The variation in retirement for the COVID-19 period arises due to a small number of individuals who report being retired during January/February 2020 and yet report some type of employment activity at the time of the COVID-19 survey completion. This employment activity is assumed to be the dominant status over baseline retirement.

Within standard deviation for hours worked is comparable in size to the pre-pandemic value for both males and females. Given the much longer time span of the pre-pandemic period in this data set, this highlights the volatility associated with the job market during the COVID-19 period. Many individuals experienced changes to their hours of work as a result of the furlough scheme, under which they agreed to a temporary period of absence from their jobs while still being paid a substantial proportion of their wage.

As expected, for the rest of the variables there is a reduction in sample within standard deviation when transitioning from a 12-year period to an 18-month one. Despite this, the differences in within variation for the rest of the variables do not raise any concerns in the context of using a within estimator for the analysis.

APPENDIX C

The self-reported well-being variable used in the current study is constructed based on responses to 12 questions of the General Health Questionnaire (GHQ). The 12 questions are presented to individuals in the following way:

"The next questions are about how you have been feeling over the last few weeks.

1. *Have you recently been able to concentrate on whatever you are doing?*
2. *Have you recently lost much sleep over worry?*
3. *Have you recently felt that you were playing a useful part in things?*
4. *Have you recently felt capable of making decisions about things?*
5. *Have you recently felt constantly under strain?*
6. *Have you recently felt you could not overcome your difficulties?*
7. *Have you recently been able to enjoy your normal day-to-day activities?*
8. *Have you recently been able to face up to problems?*
9. *Have you recently been feeling unhappy or depressed?*
10. *Have you recently been losing confidence in yourself?*
11. *Have you recently been thinking of yourself as a worthless person?*
12. *Have you recently been feeling reasonably happy, all things considered?"*

The answers to these questions are recorded on a 4-point Likert scale which ranges from 1 to 4. The answers corresponding to the numbers on the scale change depending on whether the question is 'positive' (e.g. question 1) or 'negative' (e.g. question 2). For positive answers 1 represents *"More so than usual"* to 4 which represents *"Much less than usual"*. For negative answers 1 represents *"Not at all"* to 4 which represents *"Much more than usual"*. The responses are combined to a single number by recoding so that the Likert scale runs from 0 to 3 rather than the original 1 to 4 and afterwards summing across the 12 responses. The resulting measure runs from 0 which represents "*Least distressed"* to 36 which represents *"Most distressed"*. The scale is reversed by multiplying every value by -1 and adding 36 so that higher values indicate higher levels of mental health.

APPENDIX D

Table D.1: Within estimator with one structural break for years 2009-2021 accounting for COVID-19 symptoms.

Variable	Males	Females
Time period identifier (Default: 2017)		
2009	0.147	0.401***
	(0.103)	(0.0978)
2010	0.0398	0.251***
	(0.0783)	(0.0735)
2011	0.186**	0.230***
	(0.0727)	(0.0715)
2012	0.294***	0.237***
	(0.0707)	(0.0689)
2013	0.145**	0.186***
	(0.0711)	(0.0686)
2014	0.319***	0.335***
	(0.0662)	(0.0659)
2015	0.421***	0.341***
	(0.0622)	(0.0638)
2016	0.177***	0.201***
	(0.0590)	(0.0584)
2018	-0.0468	-0.0888
	(0.0584)	(0.0574)
2019	-0.136	-0.152*
	(0.0831)	(0.0781)
January 2020	-0.231	-0.278
	(0.446)	(0.424)
February 2020	-1.082**	-0.442
	(0.499)	(0.532)
March 2020	-0.0468	0.225
	(0.735)	(1.417)
April 2020	-0.518*	-1.349***
	(0.300)	(0.267)
May 2020	-0.800***	-1.145***
	(0.307)	(0.272)
June 2020	-0.997***	-1.120***
	(0.310)	(0.274)
July 2020	-0.481	-0.302
	(0.310)	(0.275)
September 2020	-0.419	-0.540*
	(0.312)	(0.277)
November 2020	-1.155***	-1.579***
	(0.313)	(0.276)
January 2021	-1.308***	-1.586***
	(0.317)	(0.278)
March 2021	-0.869***	-0.946***
	(0.315)	(0.276)
September 2021	-0.485	-0.488*
	(0.320)	(0.278)

Not living with a partner (Default: No)		
Yes	-0.134	0.00798
	(0.119)	(0.0999)
Yes (during COVID-19)	-0.596***	-0.211*
	(0.147)	(0.114)
Child under 15 in the household (Default: No)		
Yes	-0.253***	-0.0321
	(0.0860)	(0.0790)
Yes (during COVID-19)	-0.421***	-0.488***
	(0.120)	(0.113)
Employed (Default: No)		
Yes	1.161***	0.760***
	(0.140)	(0.103)
Retired	1.738***	1.580***
	(0.136)	(0.116)
Yes (during COVID-19)	-0.815***	-0.880***
	(0.255)	(0.186)
Retired (during COVID-19)	-0.604**	-0.910***
	(0.260)	(0.240)
Housing tenure (Default: Owned)		
Owned (mortgage)	-0.340***	-0.299***
	(0.0810)	(0.0788)
Rented	-0.288**	-0.0332
	(0.139)	(0.121)
Other	-0.268	-0.128
	(0.329)	(0.308)
Government Office Region (Default: North East)		
North West	1.450**	1.281*
	(0.673)	(0.745)
Yorkshire and The Humber	1.121**	0.533
	(0.537)	(0.781)
East Midlands	0.885	0.592
	(0.584)	(0.734)
West Midlands	1.665**	0.704
	(0.695)	(0.766)
East of England	1.145**	0.745
	(0.528)	(0.751)
London	0.587	0.762
	(0.558)	(0.741)
South East	1.550***	0.841
	(0.551)	(0.738)
South West	1.472**	0.744
	(0.604)	(0.771)
Wales	1.839***	0.674
	(0.651)	(0.858)
Scotland	1.251*	0.472
	(0.760)	(0.826)
Northern Ireland	1.734	2.129*
	(1.185)	(1.100)

Long-standing health condition (Default: No)		
Yes	-0.322***	-0.360***
	(0.0807)	(0.0781)
Yes (during COVID-19)	0.346***	0.431***
	(0.113)	(0.108)
Symptoms that could be COVID-19 (Default: No)		
Yes	-0.425***	-0.529***
	(0.126)	(0.113)
Hours worked per week	-0.00476**	-0.00679***
	(0.00219)	(0.00227)
Hours worked per week (during COVID-19)	0.0141***	0.0152***
	(0.00333)	(0.00331)
Absolute income	0.0515***	0.0555***
	(0.0141)	(0.0132)
Absolute income (during COVID-19)	0.0367	0.0207
	(0.0292)	(0.0279)
Constant	23.39***	22.94***
	(0.496)	(0.680)
Observations	85,031	119,270
R-squared	0.026	0.033
AIC	455,355	675,698
BIC	455,832	676,192
Number of individuals	7,343	10,113

*Notes: Clustered-robust standard errors in parentheses; *p-value < 0.1, **p-value < 0.05, ***p-value < 0.01. R-squared refers to the within R-squared as reported by Stata. The structural break in the coefficients is generated by interacting the variables for which the note '(during COVID-19)' is reported with a binary variable which distinguishes the time after March 2020 from the previous period. For the variables capturing housing tenure and region no structural break is assumed to occur.*

Table D.2: Aggregate effect of each variable during the pandemic era based on the estimation in Table D.1.

Variable	Males	Females
	P-value	
Not living with a partner	-0.730***	-0.203
Child under 15 in the household	-0.674***	-0.520***
Employed:		
Yes	0.346	-0.120
Retired	1.134***	0.670***
Long-standing health condition	0.024	0.071
Hours worked per week	0.00934***	0.00841***
Absolute income	0.0882***	0.0762***

*Notes: *p-value < 0.1, **p-value < 0.05, ***p-value < 0.01. The associated test is a test for the significance of the aggregate effect of each variable during the pandemic era. Test is carried out using a two-sided Wald test with a null hypothesis stating that the sum of the coefficients associated with a variable in the pre-COVID-19 period and the structural break during the pandemic is equal to 0.*

APPENDIX E

Table E.1: Within estimators for job-related variables for years 2017-2021.

	1		2	
Variable	**Males**	**Females**	**Males**	**Females**
Model 1				
Employed (Default: No)				
Yes	1.362***	0.860***	1.415***	0.852***
	(0.262)	(0.188)	(0.285)	(0.205)
Retired	1.635***	1.680***	1.532***	1.766***
	(0.288)	(0.240)	(0.302)	(0.257)
Yes (during COVID-19)	-1.268***	-0.623***	-1.164***	-0.682***
	(0.308)	(0.214)	(0.332)	(0.233)
Retired (during COVID-19)	-1.074***	-1.428***	-0.801**	-1.248***
	(0.294)	(0.254)	(0.316)	(0.273)
Model 2				
Employed (Default: No)				
Yes	1.354***	0.860***	1.408***	0.854***
	(0.262)	(0.188)	(0.284)	(0.205)
Retired	1.628***	1.677***	1.526***	1.765***
	(0.288)	(0.240)	(0.302)	(0.257)
Yes (during COVID-19)	-1.301***	-0.689***	-1.182***	-0.744***
	(0.313)	(0.217)	(0.338)	(0.237)
Retired (during COVID-19)	-1.076***	-1.423***	-0.803**	-1.244***
	(0.294)	(0.254)	(0.316)	(0.273)
Furloughed during COVID-19 (Default: No)				
Furloughed	0.311**	0.283**	0.236	0.198
	(0.138)	(0.134)	(0.147)	(0.145)
Missing furlough state	-0.0413	0.0953	-0.0495	0.121
	(0.0991)	(0.103)	(0.106)	(0.111)
Model 3				
Employed (Default: No)				
Yes	1.361***	0.859***	1.411***	0.854***
	(0.262)	(0.188)	(0.284)	(0.205)
Retired	1.634***	1.681***	1.527***	1.757***
	(0.288)	(0.240)	(0.302)	(0.257)
Yes (during COVID-19)	-1.247***	-0.639***	-1.108***	-0.659***
	(0.310)	(0.216)	(0.334)	(0.235)
Retired (during COVID-19)	-1.074***	-1.432***	-0.801**	-1.247***
	(0.294)	(0.254)	(0.316)	(0.273)
Worked from home during COVID-19 if employed (Default: Never worked from home)				
Always worked from home	-0.0641	0.00253	-0.180	-0.168
	(0.105)	(0.0955)	(0.114)	(0.103)
Often worked from home	-0.0945	0.200*	-0.133	0.0682
	(0.120)	(0.111)	(0.128)	(0.118)
Sometimes worked from home	0.0167	0.0754	-0.006	0.0587
	(0.102)	(0.101)	(0.108)	(0.108)
Missing work from home status	1.589	-2.289	1.420	-2.518
	(1.140)	(1.586)	(1.369)	(1.768)

Model 4				
Employed (Default: No)				
Yes	1.367***	0.862***	1.416***	0.855***
	(0.262)	(0.188)	(0.285)	(0.205)
Retired	1.643***	1.681***	1.538***	1.767***
	(0.288)	(0.240)	(0.302)	(0.257)
Yes (during COVID-19)	-1.160***	-0.465**	-1.007***	-0.494**
	(0.309)	(0.217)	(0.333)	(0.236)
Retired (during COVID-19)	-1.083***	-1.416***	-0.807**	-1.234***
	(0.294)	(0.255)	(0.316)	(0.274)
Self-isolated during COVID-19 if employed (Default: Did not self-isolate)				
Self-isolated	-1.427***	-1.339***	-1.631***	-1.575***
	(0.276)	(0.278)	(0.280)	(0.293)
Missing self-isolating status	-0.0764	-0.174***	-0.149**	-0.209***
	(0.0677)	(0.0668)	(0.0721)	(0.0704)

*Notes: Clustered-robust standard errors in parentheses; *p-value < 0.1, **p-value < 0.05, ***p-value < 0.01. Estimation 1 incorporates the loneliness variable, whereas estimation 2 doesn't. The specified model for each estimation is not reported and is the same as in Table 1 for estimations 3 and 4 respectively, apart from the job-related variables. In order to avoid dropping further observations for the COVID-19 version of Understanding Society, an additional dummy variable for missing values is added in all estimations for all new job-related variables.*

Table E.2: Aggregate effect of each variable during the pandemic era based on the estimation in Table E.1.

	1		2	
Variable	**Males**	**Females**	**Males**	**Females**
Model 1 **(Default: Unemployed)**				
Yes	0.094	0.237	0.251	0.170
Retired	0.561**	0.252	0.731**	0.518*
Model 2 **(Default: Unemployed)**				
Employed and not furloughed	0.053	0.171	0.226	0.110
Employed and furloughed	0.364	0.454**	0.462*	0.308
Employed and missing furlough status	0.012	0.266	0.177	0.231
Retired	0.552**	0.254	0.723**	0.521*
Model 3 **(Default: Unemployed)**				
Employed and always works from home	0.050	0.223	0.123	0.027
Employed and often works from home	0.020	0.420**	0.170	0.263
Employed and sometimes works from home	0.131	0.259*	0.297	0.254
Employed and never works from home	0.114	0.220	0.303	0.195
Employed and missing work from home status	1.703	-2.069	1.723	-2.323
Retired	0.560**	0.249	0.726**	0.510*
Model 4 **(Default: Unemployed)**				
Employed and not self-isolating	0.207	0.397**	0.409*	0.361**
Employed and self-isolating	-1.220***	-0.942***	-1.222***	-1.214***
Employed and missing self-isolating status	0.131	0.223	0.260	0.152
Retired	0.560**	0.265	0.731**	0.533*

*Notes: *p-value < 0.1, **p-value < 0.05, ***p-value < 0.01. The associated test is a test for the significance of the aggregate effect of each variable during the pandemic era. Test is carried out using a two-sided Wald test with a null hypothesis stating that the sum of the coefficients associated with a variable in the pre-COVID-19 period and during the pandemic is equal to 0.*

APPENDIX F

Table F.1: Within estimator with two structural breaks for years 2009-2021.

	1		2	
Variable	**Males**	**Females**	**Males**	**Females**
Not living with partner (Default: No)				
Yes	-0.133	0.0155	0.392**	0.249
	(0.119)	(0.100)	(0.191)	(0.163)
Yes (Early COVID-19 period)	-0.631***	-0.132	-0.0648	0.186
	(0.162)	(0.127)	(0.178)	(0.137)
Yes (Late COVID-19 period)	-0.522***	-0.306**	-0.209	-0.110
	(0.175)	(0.134)	(0.167)	(0.133)
Child under 15 in the household (Default: No)				
Yes	-0.252***	-0.0285	-0.126	0.259*
	(0.0861)	(0.0790)	(0.164)	(0.143)
Yes (Early COVID-19 period)	-0.410***	-0.535***	-0.159	-0.413***
	(0.130)	(0.128)	(0.146)	(0.134)
Yes (Late COVID-19 period)	-0.457***	-0.464***	-0.0805	-0.205*
	(0.148)	(0.128)	(0.140)	(0.124)
How often feels lonely (Default: Hardly ever or never)				
Some of the time	-	-	-1.569***	-2.078***
	-	-	(0.121)	(0.0980)
Often	-	-	-5.283***	-5.267***
	-	-	(0.330)	(0.210)
Some of the time (Early COVID-19 period)	-	-	-1.483***	-1.290***
	-	-	(0.157)	(0.126)
Often (Early COVID-19 period)	-	-	-1.922***	-2.269***
	-	-	(0.438)	(0.291)
Some of the time (Late COVID-19 period)	-	-	-0.837***	-0.542***
	-	-	(0.144)	(0.119)
Often (Late COVID-19 period)	-	-	-1.981***	-2.631***
	-	-	(0.451)	(0.291)
Long-standing health condition (Default: No)				
Yes	-0.322***	-0.358***	-0.237**	-0.0435
	(0.0807)	(0.0781)	(0.118)	(0.113)
Yes (Early COVID-19 period)	0.259**	0.398***	0.123	0.237*
	(0.122)	(0.120)	(0.137)	(0.126)
Yes (Late COVID-19 period)	0.492***	0.463***	0.434***	0.371***
	(0.136)	(0.123)	(0.129)	(0.121)
Employed (Default: No)				
Yes	1.159***	0.754***	1.373***	0.867***
	(0.140)	(0.103)	(0.263)	(0.188)
Retired	1.737***	1.576***	1.637***	1.678***
	(0.137)	(0.116)	(0.288)	(0.240)
Yes (Early COVID-19 period)	-0.600**	-0.671***	-1.083***	-0.374
	(0.270)	(0.211)	(0.330)	(0.237)
Retired (Early COVID-19 period)	-0.386	-0.787***	-0.952***	-1.381***
	(0.276)	(0.258)	(0.318)	(0.271)
Yes (Late COVID-19 period)	-1.232***	-1.122***	-1.457***	-0.812***
	(0.360)	(0.224)	(0.349)	(0.231)

Retired (Late COVID-19 period)	-0.988***	-0.988***	-1.211***	-1.446***
	(0.370)	(0.307)	(0.346)	(0.300)
Hours worked per week	-0.0047**	-0.007***	-0.000919	-0.000837
	(0.00219)	(0.00227)	(0.00398)	(0.00378)
Hours worked per week (Early COVID-19 period)	0.014***	0.013***	0.0115**	0.00429
	(0.00353)	(0.00373)	(0.00458)	(0.00439)
Hours worked per week (Late COVID-19 period)	0.014***	0.018***	0.013***	0.0099**
	(0.00451)	(0.00404)	(0.00458)	(0.00421)
Absolute income	0.052***	0.056***	-0.00528	0.0475*
	(0.0141)	(0.0132)	(0.0261)	(0.0253)
Absolute income (Early COVID-19 period)	0.00636	-0.0235	0.100***	-0.0191
	(0.0293)	(0.0315)	(0.0335)	(0.0343)
Absolute income (Late COVID-19 period)	0.132**	0.0950**	0.151***	0.0807**
	(0.0573)	(0.0392)	(0.0442)	(0.0375)
Observations	85,031	119,270	40,769	57,446
R-squared	0.026	0.033	0.114	0.142
AIC	455,359	675,718	206,625	308,346
BIC	455,892	676,270	207,099	308,839
Number of individuals	7,343	10,113	7,343	10,113

*Notes: Estimation 1 incorporates 10 pre-COVID-19 waves. Estimation 2 incorporates 2 pre-COVID-19 waves and the variables associated with loneliness. Clustered-robust standard errors in parentheses; *p-value < 0.1, **p-value < 0.05, ***p-value < 0.01. R-squared refers to the within R-squared as reported by Stata. The structural breaks in the coefficients are generated by interacting the variables for which the notes '(Early COVID-19 period)' or '(Late COVID-19 period)' are reported with a categorical variable which distinguishes the time before the first structural breaks, the time between the first and second structural breaks, and the period after the second structural break. Only variables with structural breaks are reported. For all estimations, the first structural break refers to March 2020. For males, the second structural break is September 2020 for estimation 1 and June 2020 for estimation 2. For females, the second structural break is July 2020 for estimation 1 and June 2020 for estimation 2.*

Table F.2: Aggregate effect of each variable during the pandemic era based on the estimation in Table F.1.

	1		2	
Variable	**Males**	**Females**	**Males**	**Females**
Not living with a partner (Early COVID-19 period)	-0.764***	-0.117	0.327*	0.435***
Not living with a partner (Late COVID-19 period)	-0.655***	-0.291**	0.183	0.139
Child under 15 in the household (Early COVID-19 period)	-0.662***	-0.564***	-0.285*	-0.154
Child under 15 in the household (Late COVID-19 period)	-0.709***	-0.493***	-0.207	0.054
How often feels lonely:				
Some of the time (Early COVID-19 period)	-	-	-3.052***	-3.368***
Often (Early COVID-19 period)	-	-	-7.205***	-7.536***
Some of the time (Late COVID-19 period)	-	-	-2.406***	-2.620***
Often (Late COVID-19 period)	-	-	-7.264***	-7.898***
Long-standing health condition (Early COVID-19 period)	-0.063	0.040	-0.114	0.194
Long-standing health condition (Late COVID-19 period)	0.170	0.105	0.197	0.328***
Employed:				
Yes (Early COVID-19 period)	0.559**	0.083	0.290	0.493***
Retired (Early COVID-19 period)	1.351***	0.789***	0.685**	0.297
Yes (Late COVID-19 period)	-0.073	-0.368*	-0.084	0.055
Retired (Late COVID-19 period)	0.749**	0.588**	0.426	0.232
Hours worked per week (Early COVID-19 period)	0.009***	0.006**	0.011***	0.003
Hours worked per week (Late COVID-19 period)	0.009**	0.011***	0.012***	0.009***
Absolute income (Early COVID-19 period)	0.058**	0.033	0.095***	0.028
Absolute income (Late COVID-19 period)	0.184***	0.151***	0.146***	0.128***

*Notes: *p-value < 0.1, **p-value < 0.05, ***p-value < 0.01. The associated test is a test for the significance of the aggregate effect of each variable during the pandemic era. Test is carried out using a two-sided Wald test with a null hypothesis stating that the sum of the coefficients associated with a variable in the pre-COVID-19 period and the structural break during the pandemic is equal to 0.*

Table F.3: Test for the equality of the structural break coefficients for the first and second structural breaks based on the estimated model in Table F.1.

	1		2	
	Males	**Females**	**Males**	**Females**
Variable	**P-value**			
Not living with a partner	0.502	0.165	0.292	0.010
Child under 15 in the household	0.726	0.556	0.500	0.059
How often feels lonely:				
Some of the time	-	-	0.000	0.000
Often	-	-	0.882	0.208
Long-standing health condition	0.052	0.559	0.105	0.188
Employed:				
Yes	0.069	0.043	0.195	0.021
Retired	0.093	0.481	0.402	0.797
Hours worked per week	0.859	0.243	0.592	0.101
Absolute income	0.023	0.005	0.213	0.005

Notes: Carried out by using a two-sided Wald test.

APPENDIX G

Table G.1: Hansen's (1999) approach to estimating models with one, two, and three structural break dates.

Model	Estimated threshold	P-value
Specification 1:		
1st break	January 2020	0.000
2nd break	June 2020	0.640
Specification 2:		
1st break	December 2019	0.000
2nd break	June 2020	0.000
3rd break	September 2020	0.633

Notes: Specification 1 refers to the model without loneliness, and Specification 2 represents the model which incorporates loneliness. The estimated threshold represents the last month during which the existing regime holds. The p-values are based on the comparison of the test statistic with the distribution generated by the 300 bootstrap replications. The estimations are implemented by using the Stata command xthreg. *It should be noted that the periods of the time variable before 2020 represent entire calendar years. As such, the December 2019 break date refers to the entire calendar year of 2019.*

On first sight, it might appear to be the case that the assumed initial structural break date of March 2020 is not the best choice. However, the timing based on which the observations for the data set are collected is the source of this discrepancy. Observations collected during the first three months of 2020 belong to the last wave of the original Understanding Society survey. As such, there are only 331 observations for the first three months of 2020 as compared to e.g. 10,461 for April 2020 collected during the first wave of the COVID-19 version of the study.

In fact, the number of observations for the first three months of 2020 becomes 42 if we consider the balanced sample used in this case. As such, the estimated models which assume a structural break in the period between December 2019 and March 2020 are practically indistinguishable. The significant estimated structural break date of December 2019 or January 2020 is thus not contradictory to the original assumption of March 2020 as the structural break date.

THESIS CONCLUSION

Subjective, self-reported measures can provide information about unobserved variables, such as life satisfaction and mental health, making it feasible to incorporate them in quantitative analysis. However, there is scepticism around their use in applied research. Using data from the UK's Understanding Society survey this thesis: verifies their usefulness; examines how the well-being profile of individuals can be investigated by using alternative methodologies; and explores how the well-being determination mechanism might be altered in periods of crises such as the recent COVID-19 pandemic.

Chapter 1 tests the validity of a composite measure of well-being constructed using an ordinal life satisfaction measure and the GHQ measure of well-being as building blocks. The associations between the well-being measure and a set of biomarkers used to capture the overall state of health are examined. The hypothesis is that subjective well-being provides useful information on the well-being of individuals as captured by their objective health state.

To model the biomarkers together with the well-being measure a regular vine copula is used. Well-being is recorded on a discrete scale, but it is assumed that its underlying unobserved nature is continuous. In a similar fashion to utility in economics, how individuals choose to respond on the discrete scale should remain unchanged for any strict monotonic transformation of underlying well-being. Copulas are helpful in this scenario in that they can characterise the dependence between variables in a manner which is independent of the scale of each variable.

Chapter 1 finds evidence supporting the usefulness of subjective well-being measures in capturing the true levels of well-being from a health standpoint. The biomarkers for glycated haemoglobin, diastolic blood pressure, dehydroepiandrosterone sulphate, forced vital capacity, albumin, and high-density lipoprotein cholesterol are found to be significantly associated with self-reported well-being. The direction of each association points towards individuals exhibiting worse health conditions reporting lower levels of well-being.

The results support the use of subjective well-being measures currently used in the literature and in the following two chapters of the thesis which deal with understanding the determination of well-being. Understanding which factors influence well-being, as well as how they might interplay with each other, is of vital importance. If policies aim to maximise welfare, identifying the well-being profile of individuals based on observable characteristics can help identify the appropriate groups of people that should be targeted by the policy.

Much of the literature to date has used linear techniques such as OLS or the within estimator to identify significant associations with well-being. Valuable insights have emerged such as the U-shaped association of well-being with age, the positive impact of a good level of health on well-being, the negative effects of unemployment and being single, as well as the significant impact of social comparison of income on well-being. Chapter 2 contributes to this literature by using an alternative machine learning technique, namely the RE-EM tree by Sela and Simonoff (2012).

The RE-EM tree identifies patterns in the data without imposing any model structure *a priori*. It has two major advantages relative to the linear techniques often used. Firstly, it can identify non-linearities or interactions between variables which are significant for the determination of well-being without the need to identify them before estimation as in the case of linear models. Secondly, the RE-EM tree can choose the most relevant explanatory variables out of a set of variables, thus avoiding the need to estimate additional parameters for irrelevant variables like when using the standard techniques.

Chapter 2 takes the structure selected by the estimated RE-EM tree and uses it to estimate the equivalent linear model version through the within estimator to improve the comparability of the results with those of standard techniques. The estimated RE-EM tree, which proposes substantial degree of non-linearity for life satisfaction determination, has explanatory power comparable to the one of a linear model with the exact same input as the one supplied to the RE-EM tree estimation procedure.

One additional advantage of using the RE-EM tree is the fact that the relative importance of each explanatory variable used can be quantified. In this case, the health variable accounts for almost half the explanatory power in the estimated tree, followed by the job status variable, age, and the level of neuroticism.

The within estimator results representing the estimated tree structure are used to generate a set of predictive margins. These reflect the marginal associations of the explanatory variables with life satisfaction. Many of the findings echo famous literature findings. A lower level of health and unemployment are associated with a lower level of life satisfaction. Adding to the limited research on personality traits and well-being, the paper finds higher levels of neuroticism are associated with lower levels of life satisfaction. Furthermore, life satisfaction exhibits a mid-life nadir. People aged 25 to 60 are associated with a lower level of well-being relative to the rest of the sample.

Many of these well-being associations which are well-founded in the literature were stirred during the pandemic of COVID-19. The significant deterioration in mental health during the pandemic raised the question as to whether there was any structural change in the associations of each aspect of life with the level of mental health, which may be representative of mental health determination during periods of crisis. Chapter 3 explores this issue.

A structural break is detected in mental health determination, for males and females separately, with respect to core socioeconomic aspects in the lives of individuals, namely cohabitation, whether or not there is a school-aged child in the household, employment, loneliness feelings, health, hours worked per week, and income.

With the structural break assumed to be March 2020, the month during which the first national lockdown was imposed in the UK, there seems to be a negative association between the level of mental health and the frequency of feeling lonely which is reinforced after the onset of COVID-19 for males and females. Furthermore, the mental health premium of being employed or retired as opposed to being unemployed is found to be significantly reduced during the pandemic. There is also increased mental distress associated with having a child aged 15 or under in the household. A higher number of hours worked per week has a positive association with the level of mental health during the pandemic as opposed to the negative one before. Lastly, the negative mental health effect associated with having a long-term health condition in the pre-pandemic period is found to be linked with a structural change in the opposite direction during the pandemic.

The impact of the pandemic on social comparison concerns of income was also examined. It is found that pre-existing social comparison concerns persisted during the pandemic, albeit through different social comparison mechanisms for males and females. For males, the average income of others matters. For females, their income relative to others in the form of a ranking matters.

There is also tentative evidence of a second structural break in mental health determination, shortly after many of the UK measures were eased after the first wave of COVID-19. An interesting topic for further research is the investigation of whether the aforementioned relationships shift back to their pre-pandemic state as more data becomes available with time, or if there is some form of permanent structural change as a result of how the unprecedented circumstances have influenced the perception of individuals in modern UK.

In summary, the current thesis highlights the importance of subjective well-being measures in research and policy. Chapter 1 verifies their usefulness, and thus relevance for policy, by showing how subjective measures can be a good representation of the general level of well-being and health for individuals. Chapter 2 shows how the well-being profile of individuals can be investigated by using methodologies which are relatively new for the economics literature, confirming the existing findings and adding new insights. Lastly, chapter 3 offers evidence on how the well-being determination mechanism might be altered in periods of crises such as the recent COVID-19 pandemic.