UCD CENTRE FOR ECONOMIC RESEARCH

WORKING PAPER SERIES

2024

The Trajectory of Obesity in a Cohort of Irish Children and their Mothers: An Application of Sequence Analysis

David Madden
University College Dublin School of Economics

WP24/05

March 2024

UCD SCHOOL OF ECONOMICS
UNIVERSITY COLLEGE DUBLIN
BELFIELD
DUBLIN 4

The Trajectory of Obesity in a Cohort of Irish Children and their Mothers: An Application of Sequence Analysis

David Madden

(University College Dublin)

March 2024

Abstract: Sequence and cluster analysis is applied to measures of body mass index for mothers and children for four waves of the Growing up in Ireland longitudinal data set. Optimal matching analysis is used to construct a dissimilarity matrix to which cluster analysis is then applied. Distinct clusterings are found for both mothers and children and multinomial logit models are then used to investigate statistical association of covariates with membership of each group. Associations are found for maternal education and diet, and child exercise. There is no correlation between membership of groups for mothers and children.

Keywords: Obesity, sequence analysis, cluster analysis.

JEL Codes: I12, C33, C38.

Corresponding Author: David Madden,

School of Economics,

University College Dublin,

Belfield,

Dublin 4,

Ireland.

Phone: 353-1-7168396

E-Mail: david.madden@ucd.ie

This project was not funded by any specific grant.

I am grateful to Brendan Halpin and participants at an internal seminar in UCD for helpful comments. The usual disclaimer applies.

The Trajectory of Obesity in a Cohort of Irish Children: An Application of Sequence Analysis

1. Introduction

There has been much concern about rates of obesity and overweight among children and young adults, in Ireland and abroad. Ireland for example has seen an ongoing campaign entitled *Let's Take On Childhood Obesity, One Step at a Time*, co-ordinated between *Safe*food and the Irish Department of Health. International concern is reflected in the report from *The Lancet* (Swinburne et al, 2019). There is also evidence that, in some countries at least, child obesity rates may have plateaued (Keane et al, 2014, Abarca-Gomez et al, 2017).

Childhood obesity is a major cause for concern and has been described as the primary childhood health problem in developed countries (Ebbeling et al, 2002). It may be linked to a variety of serious conditions including cardiovascular dysfunction, type 2 diabetes, pulmonary, hepatic, renal and musculoskeletal complications. The clustering of cardiovascular risk factors, sometimes referred to as insulin resistance syndrome has been identified in children as young as 5 years of age (Young-Hyman et al, 2002). In addition, there are also likely to be adverse effects on health related quality of life and emotional states (Olds et al, 2011). Moreover, the effects of childhood obesity may last well beyond childhood. There is evidence that should obesity continue into adulthood, then there are increased risk factors for further serious conditions.[1] For example, evidence from a British cohort suggests that being overweight in childhood increased the risk of dying from ischaemic heart disease in adulthood two fold over 57 years (Gunnell et al, 1998).

Much of the literature on obesity at adult and childhood level has been cross-sectional in nature, analysing obesity at a given point in time and comparing its incidence across factors such as age, gender and socioeconomic status. The availability of high quality, longitudinal data, such as that available in child cohort studies however, enables analysis of the trajectory of obesity for children and young adults. The availability of similar quality data for the principal carers (in almost all cases the biological mother) of these children allows comparison of trajectories

[1] For a comprehensive analysis of the health effects of obesity and overweight across all ages see GBD 2015 Obesity Collaborators, 2017.

for children and carers from the same family over time.[2] Much of the analysis of the effect of maternal body mass index (BMI) on the trajectory of childhood obesity has focussed on pre-conception BMI of the mother (see the reviews by Mattson et al, 2019 and Heslehurst et al, 2019). The analysis here differs from this research in that it instead compares trajectories of mother and child over the same time period.

In this paper sequence analysis is employed to investigate the trajectory of obesity, using longitudinal data for Irish children/young adults aged 9, 13, 17 and 21, and for their mothers whose ages ranged from 26 to 50 when the child was aged 9.[3] Sequence analysis has been used to analyse life-course events, though its use in studies of obesity is less commonplace.[4] In sequence analysis it is the sequence of outcomes for each individual which is the basic unit of observation. Thus while there could be, say, four waves of observations for 10,000 individuals, this only counts as 10,000 observations, rather than 40,000 observations, as would be the case with panel or repeated cross-sectional analysis. The precise sequence is also important. Thus imagine three possible categories, 1, 2 and 3 and four time periods. In non-dynamic panel data analysis there is no distinction between an observation with a sequence of 1232 and one with a sequence of 2213. In sequence analysis however, these sequences are treated distinctly.

Sequence analysis also differs from mobility and transition analysis (for an application of these approaches to obesity data, see Madden, 2020). These analyses focus upon transitions and mobility between different periods, rather than looking at the sequence as a whole. Clearly the possible number of sequences will depend upon the number of periods of analysis and the possible outcomes for each period. If there are *t* periods of analysis and *s* possible outcomes then the total number of possible sequences is s^t. In particular, if there is a large number of periods, then the possible number of sequences can become very large indeed. Typically, most individuals will only follow a subset of these sequences and it is often the case that some sequences are not observed at all. Optimal matching analysis (OMA), which we outline in more detail below, can be used to obtain some metric of the "distance" between the sequences

[2] In nearly all cases in the data used here, the principal carer is the biological mother, so the terms "mother" and "maternal" will be employed.

[3] For simplicity we refer to the members of the GUI Child cohort as "children" even though by waves 3 and 4 they are more accurately described as young adults.

[4] For a review of sequence analysis and other methods of life course analysis see Mikolai and Lyons-Amos (2017) and the refences therein.

of different observations, and this information is contained in a dissimilarity matrix. Following this, hierarchical cluster analysis can then be applied to the dissimilarity matrix to investigate patterns within the data. If a clear number of clusters within the data can be identified (quite how to do this is discussed in more detail below) then these clusters can be used as dependent variables in multinomial analysis to investigate what observable characteristics are associated with membership of different clusters. An example of this approach is McVicar and Anyadike-Danes (2002) who examine transitions from school to work. In their example, they have six possible outcomes and six periods, and hence 6^6= 646656 possible sequences.

This is the approach followed here to analyse trajectories of obesity in our sample of Irish children and mothers. These children and mothers are observed on four occasions and there are three possible outcomes in terms of their BMI category: "normal" body weight, overweight and obese (as defined by age and gender adjusted BMI)). This implies there are 3^4=81 possible sequences, as the samples of children and mothers are analysed separately. By applying OMA, followed by hierarchical cluster analysis, both samples are partitioned into different groups. and the partitions by children and by mothers can be compared. This enables investigation of the degree to which sequences between mothers and children are correlated. For example, if such a correlation is observed then it could be interpreted as reflecting the influence of a common environment. It is also possible to analyse which observable covariates are associated with membership of particular groups and perhaps of most interest whether membership of a particular group for the mother is associated with membership of a given group for the child.

What advantage does this offer over more "traditional" analysis of obesity? Firstly, as outlined above, the unit of observation is the sequence, as opposed to individual episodes. This arguably provides a different perspective on the dynamics of obesity, as opposed to the calculation of mobility indices, where, in some instances the mobility index may be independent of the direction of mobility. Secondly, the cluster analysis enables the data to suggest different "types" or experiences of obesity dynamics which might not be revealed by standard regression analysis of obesity outcomes on covariates. Following on from this, it is also useful to examine the statistical association between membership of these groups and observable covariates.

The paper proceeds as follows. In the next section similar work on obesity in Ireland and elsewhere is reviewed. Sequence and cluster analysis are then discussed in more detail in section 3, explaining how OMA can be used to construct a dissimilarity matrix and then how

to identify clusters using this matrix. Section 4 discusses the data and present results before section 5 provides discussion and concluding comments.

2. Modelling BMI Trajectories

Previous analyses of the dynamics of childhood obesity in Ireland have employed latent growth curve analysis. For example, in the case of Ireland, McCrory et al (2019) estimated trajectories of childhood obesity from pooled GUI infant and child cohort data using latent growth curves[5]. This is a mixed hierarchical model with BMI for each child at time *t* a function of age and education levels. Age is included as a quadratic and there are also interaction terms between age and education. The analysis is also stratified by gender and maternal education. This stratification of course differs from the approach here, where the data suggests the stratification via the cluster analysis.

Another example of this type of analysis is Jabakhanji et al (2018) who analyse BMI trajectories across three waves of the GUI *Infant* cohort examining BMI at ages 9 months, 3 years and 5 years. The analysis is similar to McCrory et al except that a much wider range of covariates are employed in the model. Again, the approach in this paper differs from the analysis here, for the same reasons as listed above for the McCrory et al paper.

A feature of the above two approaches is that while individual variability is allowed for, the underlying assumption is that individuals belong to the same population, as represented by the single growth curve. An approach closer in spirit to the analysis here is provided in Mattsson et al (2021) who use growth mixture modelling, which allows researchers to identify different subgroups and estimate growth curves for each subgroup. Mattsson et al applied this approach to BMI measures for a sample of Irish children aged from birth to 5 years. Three distinct classes were identified: normal, high BMI at birth followed by growth and then decline and finally slightly above median BMI at birth rising then to very high BMI by age 5. Multinomial logit analysis is then carried out to investigate the association between class membership and various covariates. While this approach has many similarities to sequence analysis, one

[5] There are two cohorts in the GUI study. The *infant* cohort includes children born in 2008, while the *child* cohort includes children born in 1997-1998. The analysis in this paper is of the child cohort. The GUI dataset is discussed in more detail in section 4.

important difference is that growth curve modelling is applied to the continuous BMI measure, whereas sequence analysis is applied to the distinct categories of normal, overweight and obese.

In the international literature there are many applications of growth curve analysis to BMI trajectories for both younger and older children (see the review by Mattsson et al, 2019, and references therein).

Applications of sequence analysis to obesity categories are much less plentiful. Lacey et al (2017) combine sequence analysis and growth curve modelling to analyse the impact of life course sequences on BMI trajectories. Sequence analysis is used to construct ideal life-course sequences and the association between these sequences and BMI trajectories are then examined. To the best of our knowledge, however, there are no direct applications of sequence analysis to sequences of BMI categories.

One critical difference between the approaches adopted in the above mentioned papers and sequence analysis is the nature of the data employed. In the case of growth curve and growth mixture modelling, the data is continuous, typically raw BMI data or BMI z scores. Sequence analysis however deals with categorical data, whereby each observation in each period belongs to a discrete, mutually exclusive and mutually exhaustive category. There are advantages and disadvantages associated with both approaches. The use of continuous BMI data does provide greater granularity and arguably employs more information. However, this may come at the expense of loss of focus. Thus growth curve analysis will be sensitive to changes in BMI which may be of limited clinical importance, as for example they may involve movements within the range of the "normal" weight category, movements which may have few health consequences. In the case of sequence analysis, only movements across key BMI thresholds will "count" as a change and these movements could be regarded as being of greater importance than movements within a category.

There are also fundamental methodological differences between sequence analysis and latent growth curve analysis. The former is algorithmically based, while the latter is a modelling approach with parametric assumptions to be made by the analyst. As outlined below, sequence analysis also involves critical choices on behalf of the analyst and it is important to check for the sensitivity of results obtained to these choices.

Finally, in the literature cited above it is typically childhood trajectories which are examined on their own, or at best, the influence of pre-pregnancy maternal BMI on these trajectories. This paper differs by explicitly analysing trajectories of children and mothers together and examining the degree of association between the different clusters obtained for mothers and children following the sequence analysis.

We now turn to provide a brief account of sequence analysis and how the results can then be used to generate clusters.

3. Sequence Analysis and Cluster Analysis

Sequence analysis is a form of analysis where the fundamental observation is a sequence of "states" or categories for each individual. In its simplest case, where there is a balanced panel of individuals, there are *n* individuals, observed over *t* periods and there are *s* possible states. For each individual there will be a sequence and it is this sequence which is the key observation. The number of possible states is s^t so with higher values of *t* in particular it is clear that this number can run into thousands, though in practice it is often the case that only a subset of sequences are observed. In the application in this paper there four waves of data and hence $t=4$, and there are three possible states, "normal weight", "overweight" and "obese", and hence $s=3$ and $s^t=81$. Thus a typical sequence might be NNOvOb i.e. two states at normal weight, followed by a state overweight and then a state obese. It is important to remember that the precise ordering of states is critical here, not just the frequency of states in a given sequence.

Ultimately in most applications of sequence analysis the concern is whether individuals (children and mothers separately) with "similar" sequences can be grouped into clusters. Such an approach only makes sense if there is some means of measuring the degree of similarity (or dissimilarity) between sequences. The most common approach to measuring such dissimilarity, and which is adopted here, is Optimal Matching Analysis (OMA, Macindoe and Abbott, 2004). Dissimilarity is defined in OMA in terms of the number, order and duration of states within sequences. More specifically, given any two (different) sequences, how many operations does it take to transform one sequence into another. Operations can be insertions (adding a state to the sequence), deletions (removing a state from the sequence) or replacements (one state is replaced by another). In the case of a balanced panel, then it is replacements which

are key, since an insertion or deletion would never be required in isolation, as that would imply that sequences then become of unequal length.

Critical to the application of OMA is the specification of a cost for each operation. Costs may be defined on the basis of a priori information available to the researcher. Another approach is to use the empirical transition rates from the data to generate the substitution cost matrix, and that is the approach adopted here. A less frequently observed transition will then have a relatively higher substitution cost.

Following the application of the OMA algorithm, an *n*x*n* matrix of the dissimilarity between each observation is obtained. In practice many observations will share the same sequence and so many elements of this matrix will be zero. Cluster analysis is applied to this matrix via the *clustermat* command in Stata which applies hierarchical cluster analysis to a dissimilarity matrix. The hierarchical approach proceeds via a series of successive fusions of the sequences into groups. Thus suppose initially there are G distinct sequences in the data. The first stage of cluster analysis fuses the two most similar groups to form G-1 clusters and this process continues. At each stage of the cluster analysis sequences which are most "similar" are fused into a group, with different approaches to fusing depending upon the different ways of defining similarity/distance between groups. In this paper the Wards method is used which, at each step, finds the pair of clusters leading to the minimum within-cluster variance after merging. Initially all "clusters" have only one element and the distance between them is the squared Euclidean distance. When clusters have been formed the distance is then the squared Euclidean distance between cluster centres. The Wards method maximises the variation between clusters while minimizing that within clusters.

Ultimately, the fusion into clusters could proceed until the entire sample has been fused into one group. Hence some form of "stopping rule" is needed. There is little in the way of definitive advice on choice of stopping rule, and a mixture of statistical stopping rules and researcher's discretion is usually employed (Everitt et al, 2011). The stopping rules employed here are the Calinski-Harabasz (CH) pseudo F, the Duda-Hart (DH) index and the pseudo T squared index.

Suppose there are *n* observations in total and *k* clusters. Then the CH index is given by $CH = \frac{(TSSD - \sum_{i=1}^{k} SSD_i)/(k-1)}{(\sum_{i=1}^{k} SSD_i)/(n-k)}$ where *TSSD* is the total sum of squared distances, and SSD_i is the sum of

squared distances within group i. Effectively this compares the sum of squared distances between the clusters relative to the sum of squared distances within the clusters, adjusting for the number of clusters. If *CH* increases monotonically with *k* this is indicative of no clustering structure. If *CH* declines monotonically with *k* then this indicates a hierarchical structure while if *CH* increases to a maximum at *k* this suggests the presence of k clusters.

The other stopping rule utilized is the Duda-Hart index. Consider the case where we have k+1 and k clusters and let $DH = \frac{SSD_{k+1}}{SSD_k}$ represent the sum of squared distances for the data with k+1 clusters relative to the sum of squared distances with k clusters. Higher values of the DH statistic indicate distinct clustering, so as with the CH index, if a maximum at *k* is observed, then this suggests *k* clusters. Closely related to the DH index is the pseudo T squared index, which is the ratio of the between cluster sum of squares for *k* and *k+1* to the sum of the within cluster sum of squares of *k* and *k+1* clusters, adjusted for the number of observations in each cluster. In this case a lower value of the pseudo T squared index indicates the presence of clustering.

In common with other studies in this area subjective judgements on behalf of the researcher are also employed as the stopping rules can sometimes indicate a number of clusters which is not plausible or a number of clusters which is not helpful in terms of subsequent analysis (Everitt et al, 2011).

The next section discusses the data and present the results of the analysis.

4. Data and Results

The data in this paper comes from the first four waves of the GUI Child Cohort 98. This tracks the development of a cohort of children born in Ireland in the period November 1997-October 1998 (see Williams et al, 2009). The sampling frame of the data was the national primary school system, with 910 randomly selected schools participating in the study. Weight was measured to the nearest 0.5 kg using a medically approved flat mechanical scales and children were advised to wear light clothing. Height was measured to the nearest mm using a height measuring stick. The data also contains a wide range of information concerning the principal carers (in nearly all cases the biological mothers) of the children.

In all, the original sample in wave 1 consisted of 8568 children. Observations for where there were not valid height and weight measures were dropped, leaving a sample size from wave 1 of 8136. These children were then re-surveyed at ages 13, 17 and 21 for the second, third and fourth waves. Since we wish to follow trajectories of BMI over the four waves, we choose to use a balanced panel i.e. only those paired observations of children and mothers who appear in all waves and for whom valid height and weight measurements are available. That reduces the sample size to 4004 (2045 females and 1959 males).

In making these adjustments the issue of attrition arises. Attrition in surveys such as GUI is rarely random and this is confirmed in *A Summary Guide to Wave 4 of Growing Up in Ireland's Cohort '98 (Child Cohort) at 20 Years of Age* where it is shown that attrition tends to be higher for those with less advantaged socio-economic backgrounds. Sample weights are available, however sequence analysis does not permit the use of such weights so the results below must be interpreted in this light.

The categories, or "states" of obesity are calculated from BMI measures. BMI is obtained by dividing weight (in kilos) by height (in metres) squared. The World Health Organisation suggests a state of "underweight" for BMI from 0-18.5, "normal" for BMI between 18 and 25, "overweight" for BMI between 25 and 30, "obese" for BMI between 30 and 40 and severely obese for BMI over 40. We have very few observations with BMI either below 18.5 or above 40, so we simply use three states: normal, overweight and obese.

There is a further important issue which must be taken into account when using BMI to measure obesity in children. While the BMI thresholds for adults have general acceptance and do not differ by age or gender, the same is not true for children, where BMI can change substantially with age and gender. For example, at birth median BMI is around 13, this increases to 17 at age 1, decreases to 15.5 at age 6 and increases to 21 at age 20 (Cole et al, 2000). Cole et al (2000) provide a set of obesity/overweight cutoff points for BMI for childhood based upon international data and which they suggest should be used for international comparisons. They obtain these by drawing centile curves which pass through the adult cut-off points at age 18 and which then can be traced back to provide "equivalent" cut-off points for different ages and genders. The cutoffs are obtained by averaging data from large nationally representative surveys from Brazil, Great Britain, Hong Kong, the Netherlands, Singapore and the US, with in total nearly 200,000 observations aged from birth to 25.

The cutoffs are provided at half-yearly intervals. Thus for the first wave of data, the vast majority of children are aged 9. Assuming that age is distributed approximately uniformly within the cohort of 9 year olds, it seems appropriate to take the cut-off for age 9.5. Similarly for waves 2 and 3 of our data (who are mostly 13 and 17 year olds respectively) we use the cut-offs for ages 13.5 and 17.5. For the very small numbers of children who were aged 8 or 10 in wave 1 we use thresholds of (8.5, 10), (12.5, 14) and (16.5, 18) respectively for waves 1-3 and these cutoffs are presented in Table 1. These cutoffs have also been used in previous studies which have analysed child obesity using GUI e.g. Layte and McCrory (2011). Note that for wave 4, all observations are adults and so the WHO thresholds can be used. Since the mothers in the sample are adults for all four waves, again the standard WHO thresholds are used.

Before providing results from the sequence analysis it is useful to run through some summary statistics for the sample. Tables 2a and 2b provides rates of obesity and overweight (i.e. overweight but not obese) for the total sample of mothers and children (also by gender for the children) for the four waves. We see that for children, obesity rates increased in wave 3 and again in wave 4, especially for females (where obesity rates are generally higher across all waves). The level of overweight also moves in generally the same direction, with little change between waves 1 and 3 and then increasing in wave 4. The gender gap is less pronounced here and what seems to be happening is that for both genders BMI is increasing and children are shifting into higher categories. However, in the case of females a greater proportion are ultimately moving into the highest category.[6]

Similar trends are observed for the mothers in the sample, with the fraction obese doubling between waves 1 and wave 4. Of course, the mothers are all 11 years older by wave 4 and BMI tends to increase with age (Ogden et al, 2006 and Yeom et al, 2009).

What we are most interested in here are not just the levels of obesity and overweight, but rather the trajectory over four periods. We have three different obesity "states" which an individual can be in: normal, overweight or obese. We also have four observations for each individual: when the children are aged aged 9, 13, 17 and 20. In terms of analysis each "observation" is the sequence of states for any given individual. Thus if someone is in the normal weight

[6] We present summary statistics by gender for children. However, for the sequence and cluster analysis we present results for the children as a whole, in order to save space. Sequence and cluster analysis for the children by gender are available on request.

category when aged 9, 13 and 17 and then is overweight aged 20 their sequence reads as: N-N-N-Ov. In all there are 3^4=81 possible sequences, but clearly some are likely to be more prevalent than others and some specific sequences may not be observed at all. It should also be noted that the data is censored in the sense that we do not observe what category either the children or the mothers were in before wave 1 data was collected. Nor do we observe any transitions which may have occurred and then be reversed between data waves. Thus it is possible that a child transitioned from normal to overweight at age 10 and then back again to normal at age 12. These transitions would not be observed.

Table 3 shows the frequency table of sequences for the children and for the mothers, where we include only the top 11 sequences for brevity (sequence 10 and 11 were tied). The sequences for the children are quite concentrated, with the top 5 sequences accounting for around three quarters of all observations. Sequences are less concentrated for mothers, with the top 5 only accounting for just over half of the sequences. In other words, there is a greater diversity of sequences observed for the mothers than for the children.

The data can also be summarized graphically. Figures 1a and 1b shows the sequence index plots for children and then for mothers as first proposed by Scherer (2001). Each sequence is represented by a horizontal line, with a different colour for each state. These figures show how dominant the sequence N-N-N-N is for children, where N has the blue colour, though it is noticeable that it is less dominant for mothers.

The next stage in the analysis is to apply OMA to determine the dissimilarity matrix which will then be used in the cluster analysis and this implies finding values for the substitution costs. We choose to use a data-driven substitution cost matrix based on the transition rates between the states. For example, this process generated the following matrices for children and mothers:

$$S_c = \begin{bmatrix} 0 & 0.201 & 1.759 \\ 0.201 & 0 & 1.04 \\ 1.759 & 1.04 & 0 \end{bmatrix}$$

$$S_m = \begin{bmatrix} 0 & 0.896 & 1.535 \\ 0.896 & 0 & 0.569 \\ 1.535 & 0.569 & 0 \end{bmatrix}$$

Row 1 and column 1 refer to “normal”, row and column 2 to “overweight” and row and column 3 to “obese”. Note the matrix is symmetric as it is assumed that substitution costs are the same regardless of the direction of the transition. Thus for example, for children, the relatively low

value of 0.201 reflects the fact that the most common transitions are between normal and overweight. The high value of 1.759 reflects the fact a transition direct from normal to obese is very rare while the intermediate value of 1.040 is consistent with transitions from overweight to obesity.

The final piece of information needed to apply OMA is the indel cost. Here, indel costs are set at 75% of the highest substitution cost. This implies that an insertion plus a deletion will always have a higher cost than the corresponding substitution.[7]

OMA provides a distance or dissimilarity matrix between each sequence. Clearly for many pairwise comparisons this distance will be zero since the sequences are the same and thus the corresponding element in the matrix will be zero. Given the dissimilarity matrix, it is now possible to search for clusters, using a combination of results from the stopping rules and researchers discretion to identify the appropriate number of clusters. Tables 4a and 4b show the values for the different stopping rule indices for each cluster value for children and mothers, with the "optimal" number of clusters for each stopping rule shaded.[8] In choosing these cluster values a number of informal "rules" have been applied. First of all, the CH index simply decreases monotonically as the number of clusters increases, and thus in both cases it indicates an optimal number of k=2 clusters. The DH and pseudo T squared indices suggest optimal values in excess of two but we also apply discretion to only consider cluster solutions equal to or below five, as a number in excess of five is regarded as too unwieldy for subsequent analysis and provides too much "noise".

For the children, the CH and DH indices suggest a value of k=2, while the pseudo T squared suggests k=4. For mothers, CH suggests k=2, while DH and pseudo T squared suggest k=5.

Examination of the sequence index plots for the different clusterings on figures 2a-2d is helpful in terms of determining the characteristics of each group. Take for example, the sequence index plot for children, when k=4, figure 2b. One group consists of children who are in the normal category for every wave. A second group consists of children who move between normal and overweight, though most are overweight by wave 4. The other two groups consist

[7] I am grateful to Brendan Halpin for discussion on this.

[8] By "optimal" is meant the number of clusters which the stopping rule suggests shows the most distinct clustering pattern.

of a group who vary between overweight and obese, though by wave 4 most are obese, while the final group consists of children who have spent at least three of the four waves obese. Turning now to figure 2a, where k=2, we can see that the first two groups have been combined into a single group who have never been in the obese category, indeed many of them have been in the normal category for all four waves. Meanwhile the latter two groups (from the k=4 clustering) have now been combined into a group of children who have been obese at some stage of the four waves.

Further information regarding how the groups are formed can be obtained from examination of the dendograms. To economise on space, only the dendogram for children is presented. The dendogram shows the hierarchical clustering in the data, where the horizontal distance between each cluster essentially shows the degree of dissimilarity between them. In the case of hierarchical clustering, we start with as many clusters as observations and then keep combining groups which are "close" to each other. Figure 3 shows the latter stages of this process as we move from ten to two clusters. As a rule of thumb, the stopping rule is suggested as wherever a large horizontal gap appears on the dendogram. In figure 3 the gaps for four and two clusterings (the optimal values as suggested by the stopping rules) are clearly visible.[9]

We will now summarise the findings from the cluster analysis. For the children, we presented the sequence index plots for k=2 and 4 and we have already discussed the qualitative difference between the two clusters. As to which is the "better" or more informative clustering, much depends upon the level of granularity for which the analyst is seeking. The judgement of this researcher is that the k=4 split does provide valuable extra information.

For the mothers, stopping rules suggest an optimal value of k=2 or 5. In the case of two groups, one consists of mothers who have never been in the normal weight category for the periods we observe them. The other group, the larger group, consists of mothers who have been normal for at least one period. The five group clustering is arguably the most complex split and interestingly the groups are more homogenous in size than is the case with the clusterings for the children. There is a group where mothers are in the normal category in all four waves and there is also a group where mothers move between different categories but finish wave 4 in the normal category. There are then two groups where mothers vary between overweight and

[9] Dendograms for the other clusterings are available on request.

obese and the principal distinguishing characteristic here seems to be that for one of these groups, either the mothers predominantly spend time in the obese category, or else, in a small number of cases spend either wave 3 or 4 obese. The final group, the largest one in absolute size, consists of mothers who have all spent at least one period in the normal category, but who, by wave 4, are either overweight or obese. Again, we have taken the judgement that five is the upper limit for clustering. It is arguable that the five group clustering provides too much granularity. However, arguably the two group clustering provides too little granularity. For the subsequent analysis we will examine the groupings of k=2,4 for children and k=2,5 for mothers. A summary of the features of these clustering groups for children and mothers is provided in tables 5a and 5b.

Given that sequence data for children and their mothers are available, it is also worth investigating if there is any correlation between the clusterings i.e.in terms of the cluster analysis, is there any correspondence between the groups to which mothers and their children are assigned. This can be investigated via the Adjusted Rand Index (ARI) and Cohen's kappa. Both measures indicate the extent of agreement between two different partitions of data and are expressed relative to that agreement which would arise purely by chance. Tables 6a and 6b show the indices for the 2, 3, 4 and 5 cluster solutions. The ARI index shows very little correspondence between the partitions for children and mothers with all values less than 0.05 and some even negative. Cohen's kappa shows relatively higher values but in no case does it exceed 0.11.[10] It seems fair to say that in terms of sequences at least, there is very little correlation between mothers and children.

The final analysis we carry out is to see what (if any) observable characteristics are associated with membership of the different clusters. We do this in tables 7a-7c via multilogit analysis where the dependent variable is membership of a particular cluster and we do this for the cluster solutions suggested by the stopping rules. Hence values of 2 and 4 for children, and 2 and 5 for mothers. For children, the independent variables are maternal education, maternal health, maternal diet, maternal BMI status, maternal age, child gender, child exercise and child health. All of these variables are evaluated at their wave 1 level as we choose to fix on the value of

[10] Values of Cohen's kappa between 0.01 and 0.2 are regarded as indicating "…none to slight agreement" (McHugh, 2012).

these variables from one wave only and wave 1 gives the values at the youngest age for the child when diet and exercise habits etc are becoming established.

As additional right hand variables we also include mother's group membership according to the two clusterings (k=2,5). Thus we have two tables for children, tables 6a and 6b, but only one table for mothers (table 6c) as it seems less plausible that child group membership would influence mothers obesity trajectories.

We first of all discuss results for children. Table 7a deals with the two group clustering for children and the omitted group consists of children who have been either normal or overweight, while group 2 consists of children who have been obese in at least one wave. Most variables in the regression are statistically significant and the signs are in the "expected" direction. Those children who have been obese in at least one wave tend to have lower maternal education, are more likely to be female, exercise less, have mothers who smoke and who are or have been obese themselves. Perhaps most interesting is that mothers cluster group 1 i.e. mothers who were either obese/overweight over the waves has little predictive power, given that mothers wave 1 obesity status is already included. The next column has the same right hand variables except that now we use the five group clustering for mothers on the right hand side, and now mothers cluster membership is significant. Note that the excluded mothers group is group 4 i.e. those mothers who never become obese. All other mothers groupings are now significant and associated with child membership of group 2 i.e. children who were obese in at least one wave. It is noticeable however that the magnitude of the coefficients on mothers group membership are all quite similar.

Turning now to table 7b, where we have k=4 for children, the omitted child group becomes group 2 (consisting of children who are normal in all waves) and then higher numbered indicating longer periods spent obese. The pattern of results is qualitatively very similar to the case where k=2, with more pronounced effects (in the sense of coefficients which are higher in absolute value) observed as we move to higher groups. The protective effect of higher maternal education becomes clear as does the effect of the mother being obese in wave 1. In terms of the influence of mothers group membership, when we have only 2 groups for the mothers, a significant effect is only observed for children's membership of group 3, where children spend the final two waves as obese, but which is not the most "severe" group for children in terms of obesity.

In the most complex case where we have k=4 for children and k=5 for mothers, the largest coefficient is observed for the effect of mothers being in group 2 (highest incidence of obesity for mothers over the waves) on children being in group 3 (where they spend the final two waves being obese).

For the case of mothers in table 7c we have fewer explanatory variables (as we cannot include mother's weight category in wave 1 as we do with the children, and it is less plausible to include child cluster group membership). For the case of two groups the omitted group is that which spends time either overweight or obese. We see that membership of the other group, where at least one period is spent in the normal weight category is positively associated with higher education and negatively associated with dieting. For the case of five groups, the omitted group again is those mothers who spent most waves as obese or overweight. Higher education is associated with membership of groups spending more time in the normal weight category and we also see an effect for maternal diet.

In table 7a and 7c results are also presented using approaches do not take explicit account of the sequences. These are a random effects (RE) ordered logit panel model using the same right hand variables (with the exception of the variable for principal carer's diet which is only available for three waves and hence cannot be included in the panel analysis), and a simple fixed effect (FE) OLS model where the categories are treated as a cardinal variable.[11] A direct comparison of coefficients is complicated by the fact that the models from the cluster analysis are multinomial logits and interpretation of coefficients will depend upon the omitted group, and in addition the FE model is OLS rather than logit. In this case it seems easiest to compare signs of coefficients and levels of significance.

Going back to table 7a, the models from the clusters for children of both genders, the pattern of significance across coefficients shows a close correspondence between the cluster-derived multinomial logit models and the RE ordered logit. Higher levels of maternal education are associated with being in the normal weight category, as is child exercise. Having a mother who is overweight or obese is associated with being in the overweight or obese categories, and this is also the case for having a mother who smokes. The FE model on the other hand does not show such a strong correspondence. Maternal education is not significant at all, nor is

[11] A FE ordered logit model was also ran with qualitatively similar results which are available on request.

maternal smoking status. Significance levels for child health, exercise and maternal obesity/overweight status are all in line with the cluster and RE models.

For table 7c, which refers to mothers, care must be taken regarding interpretation of coefficients here as the omitted group in the ML analysis is the group consisting mostly of overweight and obese. Bearing that in mind, again we see a role for education in all models with higher education associated with being in the normal weight category and ill-health associated with membership of obese or overweight categories. When dealing with mothers rather than their children, the effect of smoking is in the opposite direction. For children, if the mother smoked this was associated with being overweight or obese but for the mothers themselves it appears to be associated with being normal. It is also interesting to note that in this case the FE model is far more in agreement with the ML and RE ordered logit models.

5. Discussion and Conclusion

This paper has applied sequence and then cluster analysis to the trajectories of obesity of a sample of Irish adolescents and their mothers over an eleven year period. The sequence analysis showed that only a limited number of the possible sequences were actually observed and that there was less diversity of experience among children than their mothers. This is mainly explained by the relatively high proportion (over 57 per cent) of children who never experienced any wave outside of the "normal" weight category. While this sequence of categories was also the most common for mothers, the proportion of mothers who remained in the normal category for all survey waves was markedly less, at 22 per cent. Overall, there was a greater variety of sequences amongst the mothers.

The subsequent cluster analysis showed signs of clustering amongst children and their mothers. In both cases two clustering solutions were suggested by the stopping rules: one was very simply consisting of just two groups. The other was more complex with four groups in the case of children, and five groups in the case of mothers. Notably, the clustering patterns for children and mothers showed very little correlation, which could be interpreted as evidence against some common unobserved family factor. For both mothers and children, the clusterings observed were plausible in the sense that each group had a commonality of experience with respect to the number of waves spent in each category, or in the pattern over time.

Finally, investigation of what factors are associated with membership of each clustering group revealed roles for maternal education, diet and exercise. The role of mothers group membership in terms of explaining child group membership was mixed, in the sense that for the broader clusterings (k=2 for both mothers and children) little association was found. For the more granular clusterings, some statistically significant association was found, but the pattern of association was difficult to interpret. It was also interesting to note that more traditional panel data analysis which did not exploit the sequences provided qualitatively similar results in terms of the effects of right hand variables.

It seems fair to ask how much extra insight sequence and cluster analysis adds over the more traditional panel data analysis. It encourages a more holistic analysis of the trajectory, in that the unit of observation is the complete sequence as opposed to panel data analysis whereby statistical associations arise from deviations from the within person mean over the sequence. Sequence analysis provides an attractive visualization of BMI trajectories via the sequence index plots and the sequence frequency tables provide a valuable summary of the differing concentration of sequences between children and mothers.

Cluster analysis partitioned the sample into distinct groups and while the clusterings with only two groups arguably did not provide much extra insight, clusterings with four and five groups for children and mothers respectively were able to pick out subtle differences based not just on the number of waves spent in different categories, but also on the pattern over time e.g. the distinction between groups 3 and 5 for mothers, where for both groups there is time spent in all three categories, yet group 5 differs in that all members of this group spend the final wave in the normal category.

In terms of the usefulness of the cluster analysis for policy purposes, much depends upon what we could regard as the "external validity" of the clustering. Is the pattern of clustering discovered here typical for a cohort of children/young adults in a developed country or is it simply unique to this sample? In this regard, application of the same analysis to the Infant cohort of GUI could be very helpful. However, the factors associated with membership of different groups seem to be more or less the same factors for which association is found using

more traditional analysis, and so to this extent, sequence/cluster analysis may offer relatively little extra insight to the policy-maker.

Of course, the clustering solutions here also have potential use as explanatory variables for other aspects of life course analysis (e.g. Lacey et al, 2017). That, however, is an issue for future research.

References:

Abarca-Gómez, L., Z. A. Abdeen, Z. A. Hamid, N. M. Abu-Rmeileh, B.Acosta-Cazares, C. Acuin, R. J. Adams et al (2017): "Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128· 9 million children, adolescents, and adults." *The Lancet* 390, no. 10113: 2627-2642.

Cole, T., M. Bellizzi, K. Flegal and W. Dietz (2000): Establishing a Standard Definition for Child Overweight and Obesity Worldwide: International Survey, *British Medical Journal*, Vol. 320, pp. 1-6.

Ebbeling, C., D. Pawlak, and D. Ludwig (2002). "Childhood obesity: public-health crisis, common sense cure." *The Lancet* vol. 360, pp. 473-482.

Everitt, B.S., Landau, S., Leese, M. & Stahl, D. (2011). *Cluster Analysis*. 5th ed. Wiley Publishing, West Sussex, UK.

GBD 2015 Obesity Collaborators (2017). "Health effects of overweight and obesity in 195 countries over 25 years." *New England Journal of Medicine* Vol. 377, pp. 13-27.

Gunnell, D., S. Frankel, K. Nanchahal, T. Peters, and G. Davey Smith (1998). "Childhood obesity and adult cardiovascular mortality: a 57-y follow-up study based on the Boyd Orr cohort." *The American Journal of Clinical Nutrition* vol. 67, pp. 1111-1118.

Heslehurst, Nicola, Rute Vieira, Zainab Akhter, Hayley Bailey, Emma Slack, Lem Ngongalah, Augustina Pemu, and Judith Rankin. "The association between maternal body mass index and child obesity: A systematic review and meta-analysis." *PLoS medicine* 16, no. 6 (2019): e1002817.

Jabakhanji, S., F. Boland, M. Ward, and R. Biesma (2018): "Body mass index changes in early childhood." *The Journal of Pediatrics* Vol. 202 pp. 106-114. https://doi.org10.1016/.j.peds.2018.06.049.

Keane, E. , **P. Kearney**, **I. Perry**, **C. Kelleher** and **J. Harrington (2014):** Trends and prevalence of overweight and obesity in primary school aged children in the Republic of Ireland from 2002-2012: a systematic review, *BMC Public Health* , Vol. 14**:** 974.

Lacey, R., Sacker, A., Bell, S. *et al.*(2017) Work-family life courses and BMI trajectories in three British birth cohorts. *Int J Obes* **41,** 332–339. https://doi.org/10.1038/ijo.2016.197.

Layte, R., and C. McCrory (2011): *Obesity and Overweight Among Nine Year Olds*. Government Publications. Dublin.

Lindberg, Louise, Pernilla Danielsson, Martina Persson, Claude Marcus, and Emilia Hagman. "Association of childhood obesity with risk of early all-cause and cause-specific mortality: A Swedish prospective cohort study." *PLoS medicine* 17, no. 3 (2020): e1003078.

MacIndoe, Heather, and Andrew Abbott. "Sequence analysis and optimal matching techniques for social science data." *Handbook of data analysis* (2004): 387-406.

Madden, David. "BMI Mobility and Obesity Transitions Among Children in Ireland." *Economics & Human Biology* (2020): 100896. https://doi.org/10.1016/j.ehb.2020.100896.

Mattsson, M., Maher, G. M., Boland, F., Fitzgerald, A. P., Murray, D. M., & Biesma, R. (2019). Group-based trajectory modelling for BMI trajectories in childhood: A systematic review. *Obesity Reviews*, *20*(7), 998-1015.**https://doi.org/10.1111/obr.12842**.

Mattsson, Molly, Deirdre M. Murray, Colin P. Hawkes, Mairead Kiely, Carol Ní Chaoimh, Fergus P. McCarthy, Regien Biesma, and Fiona Boland. "Body Mass Index Trajectories in the First 5 Years and Associated Antenatal Factors." *Frontiers in pediatrics* 9 (2021): 79. doi: 10.3389/fped.2021.622381.

McCrory, C., S. Leahy, A. Ribeiro, S. Fraga, H. Barros, M. Avendano, P. Vineis et al (2019): "Maternal educational inequalities in measured body mass index trajectories in three European countries." *Paediatric and Perinatal Epidemiology*, Vol. 33, pp. 226-237. doi: 10.1111/ppe.12552. Epub 2019 May 15.

McHugh, Mary L. "Interrater reliability: the kappa statistic." *Biochemia medica* 22, no. 3 (2012): 276-282.

McVicar, Duncan, and Michael Anyadike-Danes. "Predicting successful and unsuccessful transitions from school to work by using sequence methods." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 165, no. 2 (2002): 317-334. https://doi.org/10.1111/1467-985X.00641.

Mikolai, Júlia, and Mark Lyons-Amos. "Longitudinal methods for life course research: A comparison of sequence analysis, latent class growth models, and multi-state event history models for studying partnership transitions." *Longitudinal and Life Course Studies* 8, no. 2 (2017): 191-208. http://dx.doi.org/10.14301/llcs.v8i2.415.

Olds T, C. Maher, S. Zumin, S. Péneau, S. Lioret, K. Castetbon Bellisle, J. de Wilde J, M. Hohepa, R. Maddison, L. Lissner, A. Sjöberg, M. Zimmermann, I. Aeberli , C. Ogden, K. Flegal and C. Summerbell (2011): Evidence that the prevalence of childhood overweight is plateauing: data from nine countries, *Int J Pediatr Obes.*, Vol. 5-6, pp. 342-60.

Ogden, Cynthia L., Margaret D. Carroll, Lester R. Curtin, Margaret A. McDowell, Carolyn J. Tabak, and Katherine M. Flegal. "Prevalence of overweight and obesity in the United States, 1999-2004." *JAMA* 295, no. 13 (2006): 1549-1555.

Scherer, Stefani. "Early career patterns: A comparison of Great Britain and West Germany." *European sociological review* 17, no. 2 (2001): 119-144.

Swinburn, B. A., I. Vivica, S. Kraak, V. Atkins, P. Baker, J. Bogard, H. Brinsden et al. (2019): "The global syndemic of obesity, undernutrition, and climate change: The Lancet Commission report." *The Lancet* 393, no. 10173: 791-846.

Williams, J., S. Greene, E. Doyle et al (2009): *The Lives of 9 Year Olds: Growing Up in Ireland, National Longitudinal Study of Children (Report 1 of the Child Cohort).* Dublin; The Stationery Office.

Yeom, Jihye, Jung Ki Kim, and Eileen M. Crimmins. "Factors associated with body mass index (BMI) among older adults: A comparison study of the US, Japan, and Korea." *Hanʾguk Nonyonhak= Journal of Korea Gerontological Society* 29, no. 4 (2009): 1479.

Young-Hyman, D., D. Schlundt, L. Herman, F. De Luca, and D. Counts (2002). "Evaluation of the insulin resistance syndrome in 5-to 10-year-old overweight/obese African-American children." *Diabetes Care* vol. 24, pp. 1359-1364.

Table 1: Age and Gender Specific Cutoffs for Overweight and Obesity from Cole et al (2000)

	Male		Female	
Age	**Overweight**	**Obese**	**Overweight**	**Obese**
8.5	18.76	22.17	18.69	22.18
9.5	19.46	23.39	19.45	23.46
10.5	20.20	24.57	20.29	24.77
12.5	21.56	26.43	22.14	27.24
13.5	22.27	27.25	22.98	28.20
14.5	22.96	27.98	23.66	28.87
16.5	24.19	29.14	24.54	29.56
17.5	24.73	29.7	24.85	29.84
18.0	25.0	30.0	25.0	30.0

Table 2a: Obesity/Overweight rates by gender, waves 1-4 (standard errors in italics), N=4004.

	Overall				Male				Female			
	W1	**W2**	**W3**	**W4**	**W1**	**W2**	**W3**	**W4**	**W1**	**W2**	**W3**	**W4**
Ob	.056	.053	.069	.117	.048	.043	.056	.075	.064	.063	.081	.161
	.006	*.005*	*.006*	*.008*	*.007*	*.007*	*.007*	*.009*	*.009*	*.009*	*.009*	*.013*
Ov	.190	.193	.196	.233	.173	.180	.187	.244	.213	.211	.212	.221
	.009	*.009*	*.009*	*.009*	*.012*	*.012*	*.012*	*.014*	*.013*	*.013*	*.013*	*.013*

Table 2b: Obesity/Overweight rates for Mothers, waves 1-4 (standard errors in italics).

	W1	**W2**	**W3**	**W4**
Obese	0.194	0.256	0.313	0.403
	0.009	*0.010*	*0.011*	*0.011*
Overweight	0.333	0.346	0.346	0.285
	0.011	*0.011*	*0.011*	*0.010*

Table 3a: Frequency Table of Sequences - Children

Sequence Pattern	Frequency	Percent	Cumulative
N-N-N-N	2,301	57.47	57.47
N-N-N-Ov	283	7.07	64.54
N-N-Ov-Ov	166	4.15	68.68
Ov-N-N-N	122	3.05	71.73
Ov-Ov-Ov-Ov	117	2.92	74.65
N-N-Ov-N	72	1.80	76.45
N-Ov-N-N	69	1.72	78.17
Ov-N-N-Ov	64	1.60	79.77
Ov-Ov-N-N	64	1.60	81.37
Ob-Ob-Ob-Ob	57	1.42	82.79
N-Ov-Ov-Ov	53	1.32	84.12
Ov-N-Ov-Ov	53	1.32	84.12

Table 3b: Frequency Table of Sequences – Mothers

Sequence Pattern	Frequency	Percent	Cumulative
N-N-N-N	909	22.70	22.70
Ob-Ob-Ob-Ob	453	11.31	34.02
Ov-Ov-Ov-Ov	340	8.49	42.51
N-Ov-Ov-Ov	199	4.97	47.48
N-N-N-Ob	163	4.07	51.55
Ov-Ov-Ov-Ob	157	3.92	55.47
Ov-Ob-Ob-Ob	152	3.80	59.27
N-N-Ov-Ov	144	3.60	62.86
Ov-Ov-Ob-Ob	127	3.17	66.03
N-N-Ow-N	99	2.47	68.51
N-N-N-Ow	99	2.47	70.98

Table 4a: Stopping Rule Indices - Children

Number of Clusters	CH Pseudo F	Duda Hart	Pseudo T squared
1		0.5298	3552.18
2	3552.18	0.6966	205.58
3	2590.82	0.6004	2347.65
4	2499.09	0.6163	139.48
5	2438.98	0.4039	363.04
6	2388.36	0.7311	451.21
7	2427.37	0.6853	324.19
8	2404.2	0.5514	82.18
9	2442.19	0.5958	82.09
10	2464.79	0.2625	1458.35
11	2527.99	0.6059	290.75
12	2623.11	0.0371	2152.64
13	2709.44	0.247	265.28
14	2840.91	0.5575	46.03
15	2996.07	0.4391	328.33

Table 4b: Stopping Rule Indices – Mothers

Number of Clusters	CH Pseudo F	Duda Hart	Pseudo T squared
1		0.5844	2846.45
2	2846.45	0.6691	1203.22
3	2485.58	0.5578	1242.37
4	2245.91	0.5598	1052.18
5	2096.7	0.7137	438.42
6	2107.8	0.5868	480.87
7	2081.88	0.5341	791.31
8	2071.09	0.7076	177.29
9	2027.66	0.5478	336.79
10	2030.45	0.6226	398.88
11	2009.79	0.5259	310.15
12	2018.65	0.6924	136.39
13	2055.43	0.6436	251.39
14	2081.35	0.5356	275.72
15	2128.01	0.2682	867.51

Table 5a: Features of Cluster Groupings, k=2

	Children	Mothers
Group 1	Either normal or overweight in every wave.	Either obese or overweight in every wave
Group 2	Obese in at least one wave	Normal in at least one wave

Table 5b: Features of Cluster Groupings, k=4, 5

	Children	Mothers
Group 1	Never obese in any wave, overweight in at least one wave	Always obese or overweight. Some only overweight
Group 2	Normal in every wave	Always obese or overweight, always obese in at least one wave. Some only obese.
Group 3	Obese in at least one, but never more than two waves. Obese in wave 4	Mixture of normal, overweight and obese, but either obese or overweight in final wave.
Group 4	At least two waves spent obese. Never normal in any wave	Normal in every wave
Group 5		Mixture of normal, overweight and obese, but all normal in final wave.

Table 6a: ARI index between Clustering for Children and Mothers

Childrens Groups	**Mothers Groups**			
	k=2	**k=3**	**k=4**	**k=5**
k=2	0.0415	-0.0029	0.0163	0.0050
k=3	0.0417	-0.0051	0.0161	0.0041
k=4	0.0483	0.0064	0.0288	0.0133
k=5	0.0484	0.0057	0.0287	0.0130

Table 6b: Kappa max index between Clustering for Children and Mothers

Childrens Groups	**Mothers Groups**			
	k=2	**k=3**	**k=4**	**k=5**
k=2	0.1140	0.0629	0.0556	0.0456
k=3	0.0841	0.0429	0.0437	0.0352
k=4	0.1140	0.0739	0.0681	0.0581
k=5	0.1140	0.0760	0.0678	0.0557

Table 7a: Multinomial Logistic Regression on Cluster Membership, children (k=2)

	k=2	k=2	REOL	FE
Variables (all W1)	**Group 1**	**Group 1**		
Leaving Cert	-0.357**	-0.351**	-0.114	0.003
	(0.145)	(0.145)	(0.142)	(0.025)
Non-degree	-0.666***	-0.662***	-0.478***	-0.021
	(0.160)	(0.160)	(0.154)	(0.026)
3rd Level	-0.844***	-0.828***	-0.634***	0.013
	(0.165)	(0.165)	(0.154)	(0.027)
Maternal Age	-0.007	-0.006	-0.031***	
	(0.010)	(0.010)	(0.011)	
Gender	0.239**	0.237**	0.172	
	(0.104)	(0.104)	(0.111)	
Mum Chr Hlth	0.033	0.039	0.100	0.003
	(0.150)	(0.150)	(0.089)	(0.013)
Mum Smokes	0.552***	0.536***	0.305***	-0.006
	(0.126)	(0.126)	(0.107)	(0.016)
Mum Diet	0.304**	0.287**		
	(0.119)	(0.119)		
Child Exercise	-0.203***	-0.202***	-0.200***	-0.024***
	(0.035)	(0.035)	(0.014)	(0.002)
Child Sick	-0.118	-0.145	0.510***	0.072***
	(0.469)	(0.463)	(0.174)	(0.025)
Mother Obese	1.013***	0.816***	0.718***	0.039***
	(0.222)	(0.251)	(0.086)	(0.011)
Mother OW	0.350*	0.187	1.284***	0.085***
	(0.187)	(0.209)	(0.104)	(0.015)
Mother group 1	0.217	0.649**		
	(0.186)	(0.283)		
Mother group 2		0.736**		
		(0.288)		
Mother group 3		0.492***		
		(0.187)		
Mother group 5		0.451*		
		(0.255)		
Constant	-0.970*	-1.292**		1.378***
	(0.513)	(0.522)		(0.025)
Observations	**4004**	**4004**	**4004**	**4004**

Table 7b: Multinomial Logistic Regression on Cluster Membership, children (k=4)

	k=4	k=4	k=4	k=4	k=4	k=4
Variables (all W1)	**Group 1**	**Group 3**	**Group 4**	**Group 1**	**Group 3**	**Group 4**
Leaving Cert	0.016	-0.135	-0.557***	0.020	-0.126	-0.551***
	(0.123)	(0.197)	(0.202)	(0.123)	(0.197)	(0.203)
Non-degree	0.114	-0.383*	-0.860***	0.118	-0.373*	-0.856***
	(0.126)	(0.216)	(0.232)	(0.126)	(0.215)	(0.231)
3rd Level	-0.233*	-0.786***	-1.039***	-0.228*	-0.764***	-1.026***
	(0.126)	(0.226)	(0.233)	(0.126)	(0.225)	(0.233)
Maternal Age	-0.013*	-0.035**	0.013	-0.013*	-0.035**	0.013
	(0.007)	(0.014)	(0.015)	(0.007)	(0.014)	(0.015)
Gender	-0.031	0.448***	-0.015	-0.030	0.448***	-0.016
	(0.072)	(0.142)	(0.146)	(0.072)	(0.142)	(0.146)
Mum Chr Hlth	-0.074	-0.015	0.020	-0.071	-0.006	0.025
	(0.111)	(0.200)	(0.215)	(0.111)	(0.201)	(0.215)
Mum Smokes	0.232**	0.423**	0.883***	0.221**	0.400**	0.865***
	(0.101)	(0.168)	(0.177)	(0.101)	(0.170)	(0.177)
Mum Diet	0.438***	0.474***	0.521***	0.426***	0.449***	0.503***
	(0.094)	(0.158)	(0.172)	(0.095)	(0.159)	(0.170)
Child Exercise	-0.077***	-0.146***	-0.323***	-0.076***	-0.146***	-0.321***
	(0.028)	(0.050)	(0.049)	(0.028)	(0.050)	(0.049)
Child Sick	-0.006	0.443	-1.445	-0.030	0.397	-1.463
	(0.336)	(0.493)	(1.090)	(0.339)	(0.488)	(1.088)
Mother Obese	0.486***	0.643*	1.774***	0.415**	0.373	1.611***
	(0.166)	(0.329)	(0.299)	(0.181)	(0.378)	(0.330)
Mother OW	0.261**	-0.000	0.877***	0.180	-0.213	0.720**
	(0.125)	(0.280)	(0.253)	(0.135)	(0.315)	(0.282)
Mother group 1	0.147	0.565**	0.034	0.368**	1.066***	0.511
	(0.131)	(0.281)	(0.243)	(0.179)	(0.399)	(0.407)
Mother group 2				0.348*	1.208***	0.532
				(0.185)	(0.411)	(0.405)
Mother group 3				0.247**	0.556**	0.563*
				(0.108)	(0.239)	(0.296)
Mother group 5				0.171	0.578*	0.399
				(0.146)	(0.331)	(0.386)
Constant				-0.007	-1.169*	-1.530*
				(0.400)	(0.670)	(0.781)
Observations	**4004**	**4004**	**4004**	**4004**	**4004**	**4004**

Table 7c: Multinomial Logistic Regression on Cluster Membership, Mothers

	k=2	k=5				REOL	FE
Variables (all Wave 1)	**Group 2**	**Group 1**	**Group 2**	**Group 4**	**Group 5**		
Leaving Cert	0.163	-0.034	-0.211	0.235	-0.244	-1.311***	-0.301***
	(0.111)	(0.168)	(0.147)	(0.167)	(0.188)	(0.148)	(0.039)
Non-degree	0.240**	-0.078	-0.299*	0.267	-0.268	-1.901***	-0.427***
	(0.115)	(0.174)	(0.153)	(0.171)	(0.194)	(0.158)	(0.043)
3rd Level	0.497***	-0.106	-0.734***	0.338**	-0.326*	-2.047***	-0.411***
	(0.115)	(0.170)	(0.156)	(0.166)	(0.189)	(0.163)	(0.049)
Maternal Age	-0.004	0.016	0.011	0.018*	0.020*	-0.035***	
	(0.007)	(0.010)	(0.009)	(0.009)	(0.012)	(0.011)	
Mum Chr. Health	-0.185*	0.169	0.193	0.054	-0.117	0.383***	0.047***
	(0.101)	(0.148)	(0.137)	(0.141)	(0.183)	(0.076)	(0.017)
Mum Smokes	0.029	-0.248*	-0.131	-0.324**	-0.260	-1.048***	-0.260***
	(0.094)	(0.138)	(0.123)	(0.128)	(0.159)	(0.104)	(0.025)
Mum Diet	-1.458***	0.882***	1.404***	-0.969***	0.048		
	(0.086)	(0.125)	(0.113)	(0.172)	(0.166)		
Constant	0.051*	-0.930*	-0.712	-1.452***	-1.874***	-1.452***	2.204***
	(0.026)	(0.496)	(0.459)	(0.469)	(0.582)	(0.469)	(0.035)
Observations	**4004**	**4004**	**4004**	**4004**	**4004**	**4004**	**4004**

Figure 1a: Sequence Index Plot- Children

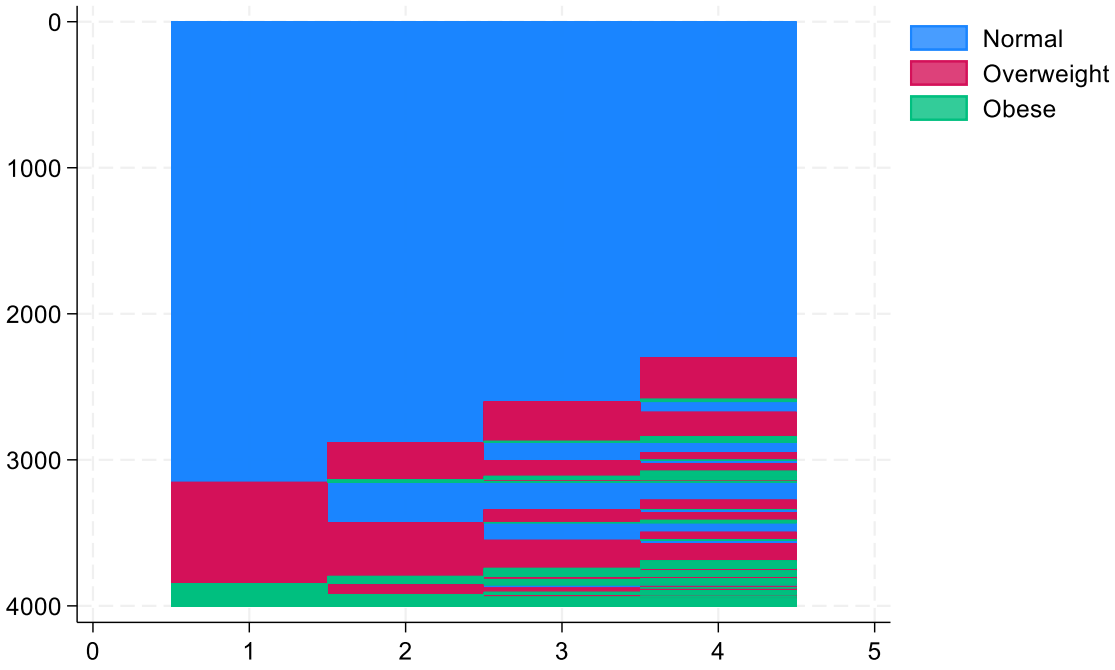


Figure 1b: Sequence Index Plot- Mothers

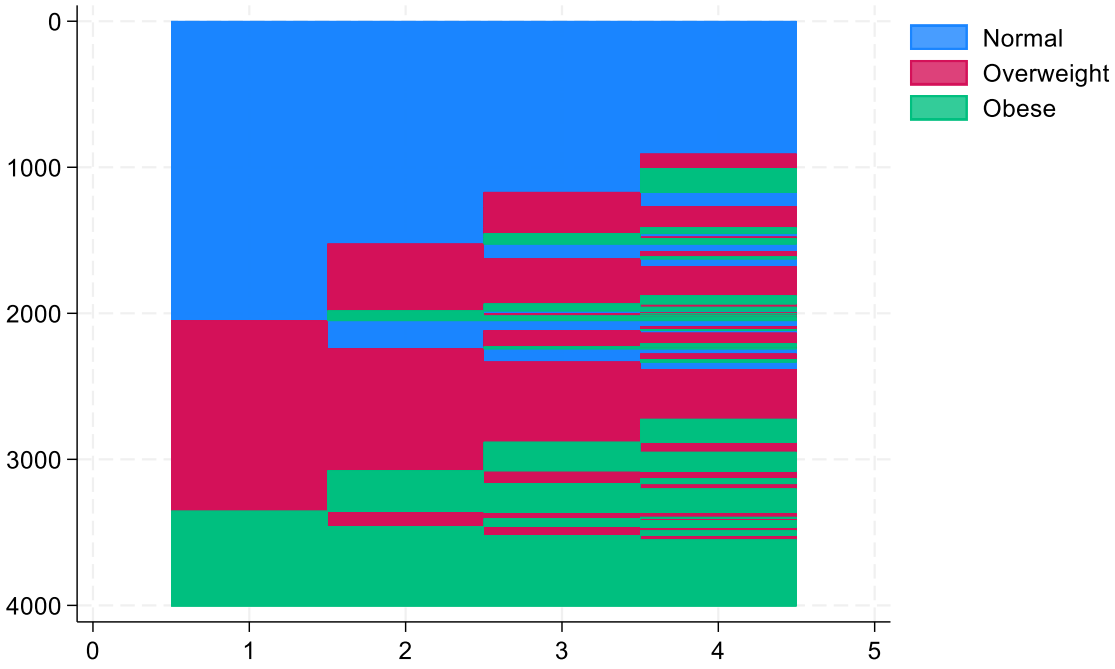


Figure 2a: Sequence Index Plots for Children, k=2

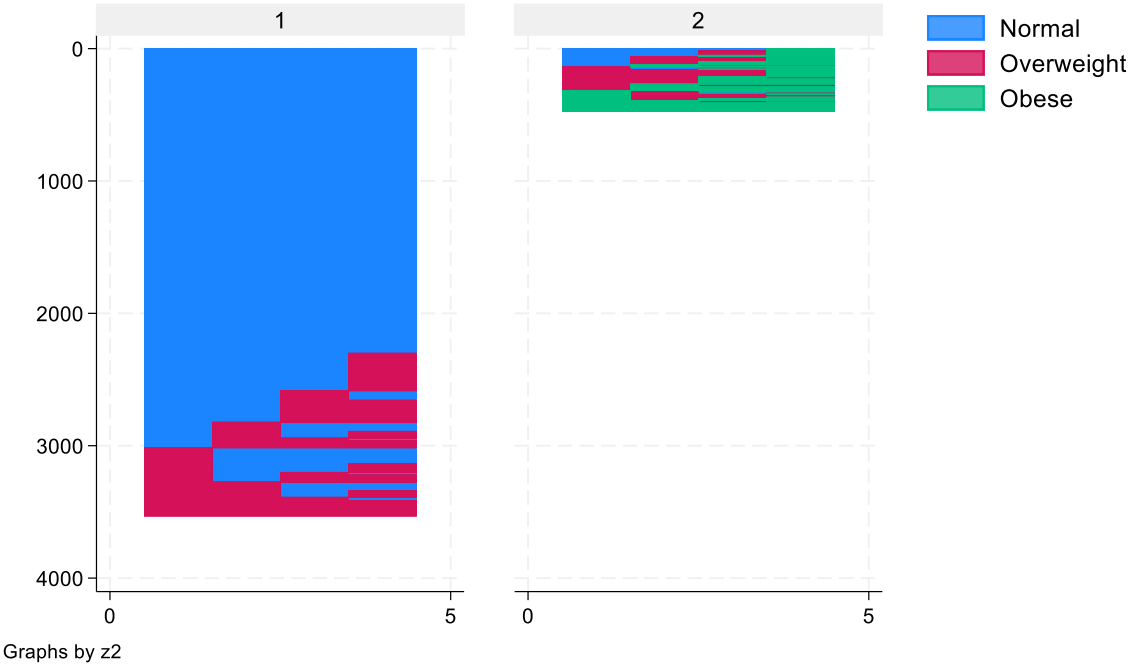


Figure 2b: Sequence Index Plots for Children, k=4

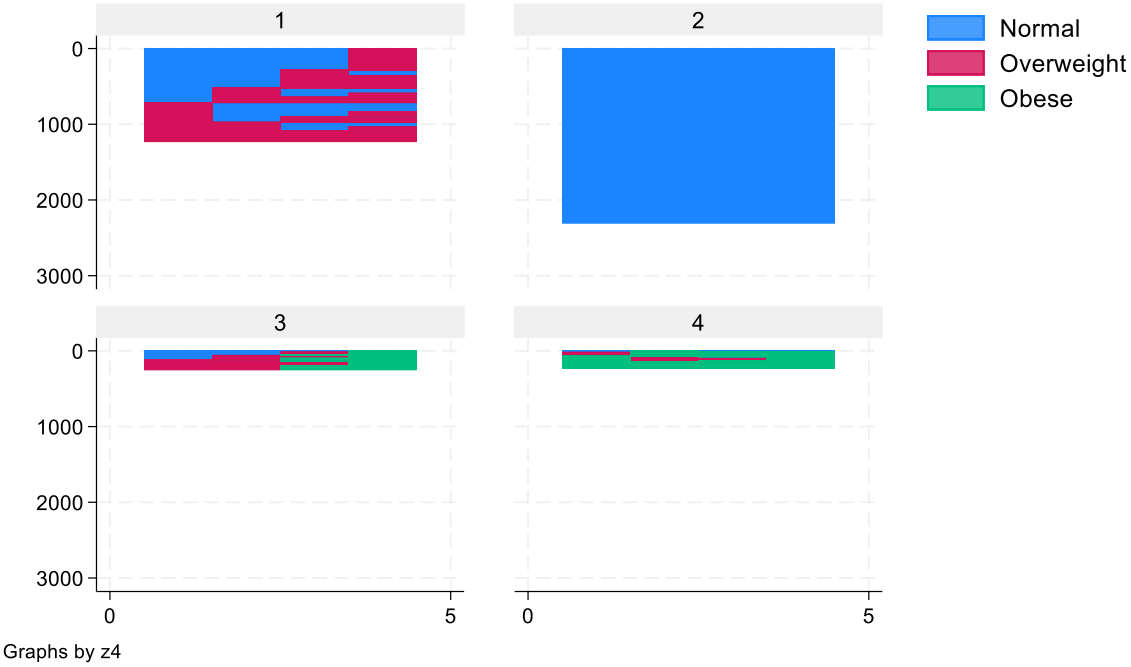
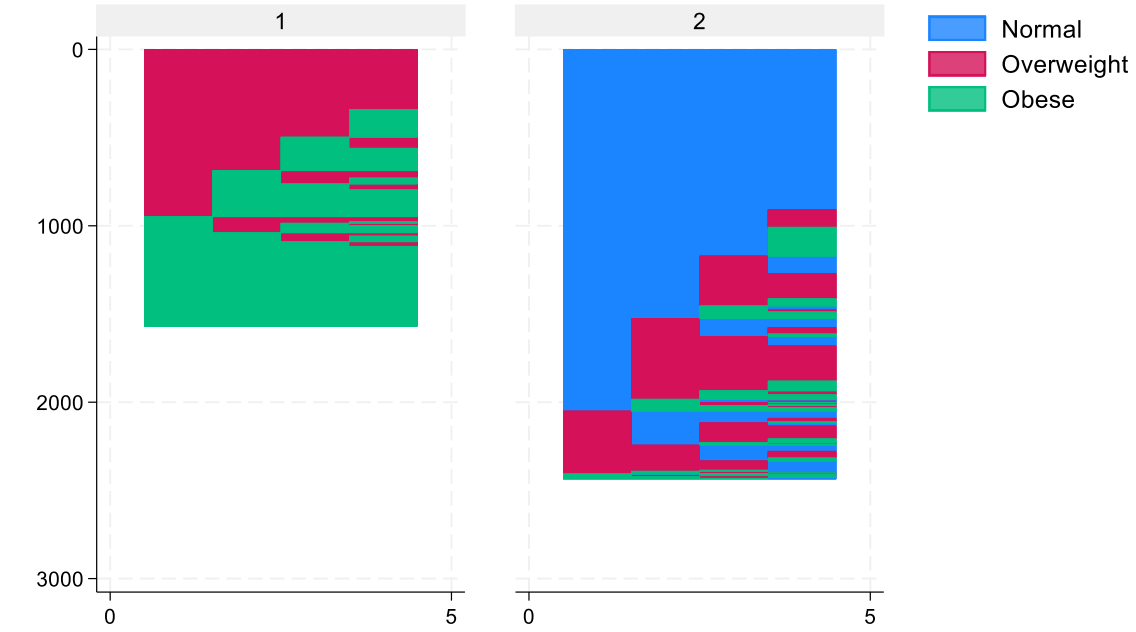
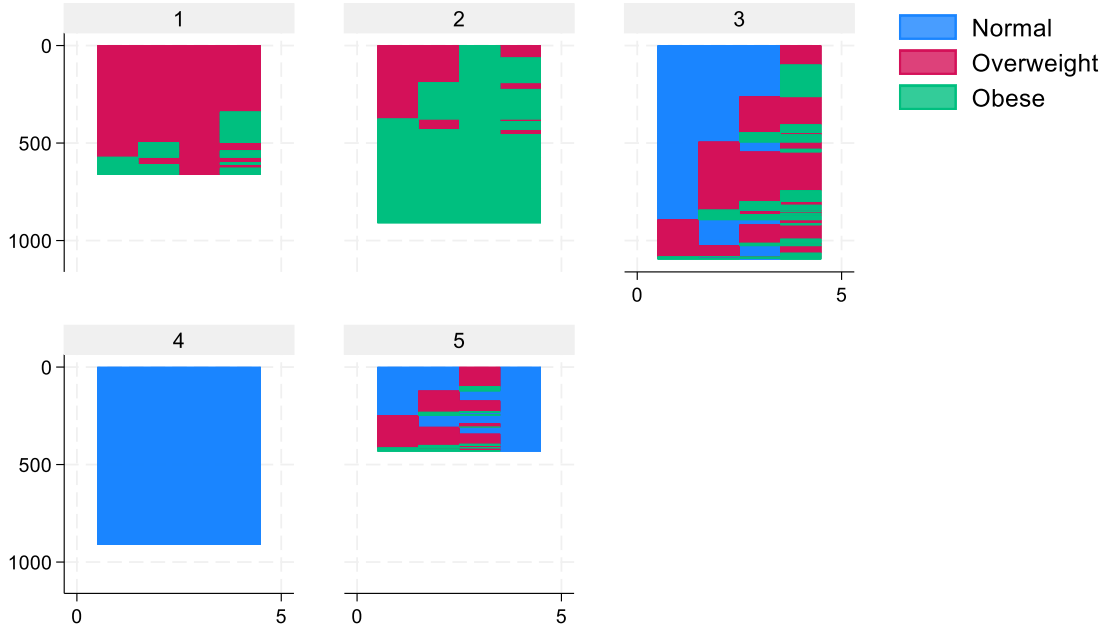


Figure 2c: Sequence Index Plots for Mothers, k=2



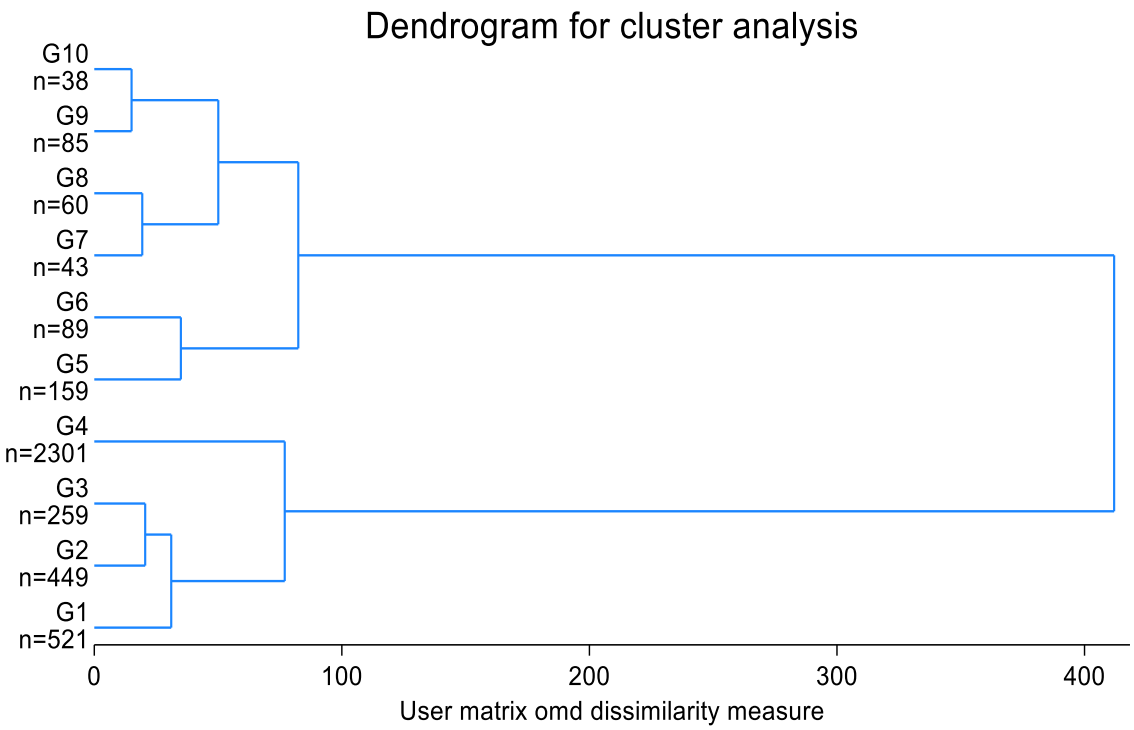
Sequence index plots for k=2

Figure 2d: Sequence Index Plots for Mothers, k=5



Sequence index plots for k=5

Figure 3: Dendogram for children



UCD CENTRE FOR ECONOMIC RESEARCH – RECENT WORKING PAPERS

WP23/10 Manuel E. Lago: '#MeToo...or not? Do salient shocks affect gender social norms?' April 2023
WP23/11 Cormac Ó Gráda, Tyler Anbinder, Simone A. Wegge: 'Gaming theSystem: The Not-So-Poor and Savings Banks in Antebellum New York' April 2023
WP23/12 Tadgh Hegarty, Karl Whelan: 'Disagreement and Market Structure in Betting Markets: Theory and Evidence from European Soccer' May 2023
WP23/13 Tadgh Hegarty, Karl Whelan: 'Do Gamblers Understand Complex Bets? Evidence From Asian Handicap Betting on Soccer' May 2023
WP23/14 Matthew Amalitinga Abagna: 'Special Economic Zones and Local Economic Activities in Ethiopia.' May 2023
WP23/15 Gregory Corcos, Stefanie Haller: 'Importer Dynamics: Do Peers Matter?' July 2023
WP23/16 Karl Whelan: 'How Does Inside Information Affect Sports Betting Odds?' July 2023
WP23/17 Michele Gubello: 'Social Trust and the Support for Universal Basic Income' July 2023
WP23/18 Constantin Bürgi, Zuzanna Studnicka: 'Online Interactions, Trade And The Gravity Model' August 2023
WP23/19 Juan S. Morales, Margaret Samahita: 'Can Social Pressure Stifle Free Speech?' August 2023
WP23/20 Fabrice Kämpfen, Xavier Gómez-Olivé, Owen O'Donnell, Carlos Riumallo Herl: 'Effectiveness of Population-Based Hypertension Screening: A Multidimensional Regression Discontinuity Design' August 2023
WP23/21 Annette Alstadsæter, Ronald Davies, Mathieu Parenti, Farid Toubal: 'The Real Effects of Tax Havens' October 2023
WP23/22 Ronald B. Davies, Dieter F. Kogler, Guohao Yang: 'Construction of a Global Knowledge Input-Output Table' October 2023
WP23/23 Taipeng Li, Lorenzo Trimachi, Rui Xie, Guohao Yang: 'The Unintended Consequences of Trade Protection on the Environment' October 2023
WP23/24 Constantin Bürgi, Mengdi Song: 'Do Professional Forecasters Believe in Uncovered Interest Rate Parity?' November 2023
WP23/25 Morgan Kelly, Kevin Hjortshøj O'Rourke: 'Industrial policy on the frontier: lessons from the first two industrial revolutions' November 2023
WP23/26 Morgan Kelly, Kevin Hjortshøj O'Rourke: 'Industrial policy on the frontier: lessons from the first two industrial revolutions' November 2023
WP23/27 Constantin Bürgi, Prachi Srivastava, Karl Whelan: 'Oil Prices and Inflation Forecasts' November 2023
WP23/28 Cormac Ó Gráda, Chihua Li, Ann Arbor, L. H. Lumey: 'How Much Schizophrenia Do Famines Cause?' December 2023
WP23/29 Judith M. Delaney, Paul J. Devereux:'Gender Differences in Teacher Judgement of Comparative Advantage' December 2023
WP23/30 Vincent Hogan, Patrick Massey: 'Different Strokes: Winning Strategies in Women's (and Men's) Big Bash Cricket.'
WP24/01 Ronald B. Davies, Lena S. Specht: 'Brexit and Foreign Students in Gravity' February 2024
WP24/02 Ronald B. Davies, Guohao Yang: 'A Comparison between Traditional and Knowledge Input Output Tables' February 2024
WP24/03 Tadgh Hegarty, Karl Whelan: 'Comparing Two Methods for Testing the Efficiency of Sports Betting Markets' February 2024
WP24/04 Matthew Amalitinga Abagna, Cecília Hornok, Alina Mulyukova: 'Place-based Policies and Household Wealth in Africa' February 2024

UCD Centre for Economic Research
Email economics@ucd.ie