Jointly modeling time-varying offending profiles and criminal career trajectories in a sample of sexual offenders

A multivariate approach to modeling criminal careers.

Mat Weldon
m.weldon@lancaster.ac.uk
Department of Mathematics and Statistics
Lancaster University

September, 2012

Total pages excluding frontmatter and references: 48.

Abstract

The study begins by summarising three key topics of research in criminal careers of sexual offenders: explaining the observed bimodality of the age-crime curve for sexual offenders; determining whether rapists and child molesters follow well-separated, or overlapping criminal career paths; and assessing whether non-contact sexual offending is associated with serious sexual crime, and if so, which serious sexual crimes it is associated with. It is argued that these and other salient questions in the study of criminal careers of sexual offenders cannot easily be answered using trajectory models that aggregate crimes across types, or profile models that account for mix of crimes but do not take account of criminal career dynamics.

Using data on 824 male sexual offenders from the Massachusetts Treatment Center, the study investigates four methods for the joint analysis of sexual offender trajectories and profiles. The four methods are: optimal matching with non-probabilistic clustering; “constrained” multivariate group based trajectory models; “unconstrained” multivariate group based trajectory models; and a Poisson-log normal factor model with posterior trajectory analysis of the factor scores.

Optimal matching is found to be a useful way to summarise and visualise complex sequences of events, with the advantage that clustering can be performed without aggregating the data. However, without the ability to make inferences from findings its usefulness is limited to an exploratory role.

The constrained and unconstrained multivariate trajectory models are compared, and it is found that neither one dominates the other in all contexts, and that both correspond to interesting theoretical hypotheses. It is suggested that the unconstrained models should be a starting point for modeling, since they make fewer assumptions and allow the appropriateness of the constrained model to be tested.

It is demonstrated that the factor model can produce parsimonious trajectories for different types of criminal activity that are close to independent at each time point, but which are nevertheless highly associated (positively or negatively) over the life course. In applying the model, it is shown that a three factor model distinguishes between trajectories characterised by general crime, rape and child molestation.

With regards to the research questions, all of the methods lend evidence to indicate the existence of bimodal trajectories, and that these trajectories exist at the individual level and are not an artifact of aggregation. However, it is shown that it is not possible to answer questions relating to associations between the occurrences of certain types of crime, using a dataset in which certain combinations of occurrence and non-occurrence are structurally missing.

Acknowledgements

I would like to thank Kelly, my wife, who has patiently supported my sometimes obsessional working patterns over the past 12 weeks, and for the last year. I would also like to thank Mae and Frankie, my children, who helped me to forget the dissertation occasionally and have fun.

I would also like to thank Professor Brian Francis, my supervisor, for his support and guidance, and for introducing me to this field. I also owe a debt of gratitude to Brian for supporting my application to study at the University and my application for funding to the North West Doctoral Training Centre (NWDTC).

Dr. Danielle Harris, of San Jose State University, California, compiled the sample used in the study, which must have taken many hours of conscientious work. I would like to thank Danielle for sharing the dataset with us, and for friendly and helpful answers to any questions I had about the data.

Thank you also to the NWDTC and the ESRC who have financially supported my studies.

Contents

1 Introduction

The Group Based Trajectory Model (GBTM) is a widely used model for the analysis of criminal careers, represented as "trajectories" - smooth curves of frequency of offending over time. One of the main advantages of GBTMs is their ability to reduce the dimensionality of complex trajectories of offending, from initially, N separate trajectories, to a small number, to enable the qualitative interpretation of patterns of offending over time. In recent years, they have increasingly been used to analyse offending trajectories of sexual offenders (Hanson, 2002; Lussier, 2010), a group of offenders who have traditionally been treated as unique and incomparable to general offenders within criminological research (Soothill et al., 2000).

The results of the application of trajectory models to sexual offending have, in some instances, confirmed that sexual offenders are similar to other types of offender, with similar trajectories of offending and a large degree of overlap between sexual and other types of offending. In other instances, studies have seemed to show ways in which sexual offending, or certain types of sexual offending, are indeed unique. For example, studies on sexual offenders have seemed to show bimodal offending trajectories, and the presence of late-onset accelerators into middle age, neither of which are predicted by the classic theoretical models (c.f. Moffitt (1993); Gottfredson and Hirschi (1990)). Other studies have also cast doubt on the idea that sexual offending is a homogenous and well-separated phenomenon, by showing that sexual offenders both engage in other types of non-sexual crime *and* specialise in certain types of sexual offending, sometimes at the same time (Soothill et al., 2000). If sexual offenders are not a homogenous and well-separated group, with regards to offending profiles as well as trajectories, then it is necessary to use methods that can discover this hidden heterogeneity in the cross-sectional plane (profiles of criminal offending at any one time) as well as the longitudinal (trajectories of offending over time).

The extant questions in criminal careers research for sexual offending: questions of sequencing, escalation in crime seriousness and versatility/specialisation - are not usually addressed using GBTMs, in which crimes of different types tend to be aggregated into a single frequency of offending by age. Other types of model exist, that are well suited to modeling criminal profiles. Latent class analysis, sometimes called latent profile analysis, is simply the cross-sectional counterpart to a group based trajectory model. These models have been used to analyse profiles of sexual burglary by Pedneault et al. (2012). Elsewhere, correspondence analysis has been used as a way to represent the strength of categorical association between categories of sexual and general offence (Soothill and Francis, 1999). Longitudinal models for changing profiles of offending include latent markov models (Bartolucci et al., 2007). However, none of these models take account of changing frequency of offending over time, or of the interaction between changing frequency and changing offending profiles.

Using data on 824 male sexual offenders from the Massachusetts Treatment Center, this study will investigate interactions between trajectories and profiles of offending directly, by employing models that extend the GBTM to deal with multivariate data. In addition to employing existing models for multivariate trajectories, we will consider a method that makes use of factor analysis to reduce the dimensionality of criminal profiles before analysing trajectories of the resulting factor scores. As far as we know this method has not been used before in the context of criminal careers research. The method will be evaluated based on its usefulness for jointly modeling trajectories and criminal profiles.

We will also show how non-probabilistic data mining approaches can be employed to aid in the exploration of complex sequences of crime occurrence and assist the formation of hypotheses. In so doing, we will argue that, despite their limitations, sequence mining methods are a powerful complement to model-based methods.

In the next section we will outline the development of GBTMs within criminal careers research for general and sexual offending, and we will trace the background to some current issues in the study of sexual offending trajectories.

1.a Background to criminal careers research

An important concept in criminology is the age-crime curve. This concept can be used to describe the changing prevalence and also the changing incidence, of crime with age. At an aggregated level, the shape of the age-crime curve has been found to be remarkably stable across groups, crime types and contexts (Hirschi and Gottfredson, 1983). The typical age-crime curve describes the occurrence of crime rising rapidly throughout adolescence, peaking before 20 and declining sharply thereafter, with a fat tail. However, what has been less clear is the extent to which the regularity of the age crime curve is caused by changing prevalence with age (i.e. the changing proportion of people participating in crime) or changing incidence with age (i.e. a change in frequency of commission of crimes for individuals) (Farrington, 1986).

In this context, in the early 1990's the emphasis of analyses of the age-crime curve shifted from the aggregate level to the individual longitudinal level. This change of emphasis was driven by the development of two, somewhat incompatible, theoretical models for criminal careers. On the one hand, Moffitt (1993) argued that the age crime curve was decomposable into two distinct groups, termed "adolescent-limiteds" and "life-course persistents", with age-crime curves (trajectories) that were different not just in magnitude, but also in shape. On the other hand, Gottfredson and Hirschi (1990) asserted that the level of criminal activity of an individual was a function only of a single time-stable underlying factor, or propensity, that was related to lack of self-control. This theory allowed for variation of criminal activity with age, but not the interaction between propensity and age, with the implication that criminal trajectories for all people should be proportional and have the same shape.

The inconsistency between these two theories was clear, and new methods were needed that could test the validity of both. In this context, Nagin and Land (1993) introduced a new method for clustering individual criminal careers, based on finite mixture models, that facilitated the testing of such theories. Together, this method and theory have spawned a multitude of studies devoted to teasing out "latent trajectories" of criminal careers.

The group based trajectory or latent trajectory model, as it has become known, has been extended and generalised to allow the analysis of multivariate trajectories. This has mainly focused on modeling variables that are disjoint in time, for example, the co-modeling of pre-adolescent indicators of anti-social behaviour and later involvement in crime. However, they have tended to be limited to the analysis of at most two variables, and the analysis of associations between variables has tended to be secondary to the classification of individuals.

One area that has, until relatively recently, not been subjected to group based trajectory analysis has been the area of sexual offending. The study of sexual offending has tended to be a specialised area of criminology with its own theories and methods. This has been at least partly due to an implicit assumption that sexual offenders are completely distinct from, and incomparable to, general offenders. This assumption is counter to the criminal propensity theory of Gottfredson and Hirschi, which would predict sexual offending to be caused by the same propensities as general offending. Much of the work that has bridged the gap between sexual offending research and criminal careers research in recent years has been concerned with examining this assumption, asking in what ways sexual offending is unique, and in what ways is it similar to and related to other types of offending.

Soothill et al. (2000) showed that sexual offenders were actually likely to be involved in other non-sexual offending throughout their lives. They also demonstrated, somewhat counterintuitively, that sexual offenders often specialised in one type of sexual offending, whilst remaining versatile in their non-sexual offending. Using the same sample as this study, Harris et al. (2009) analysed specialisation by using specialisation and diversity indices, and found that the male sexual offenders in the sample were versatile on the whole, although child molesters were more likely to specialise than rapists.

Studies have also addressed the question of whether trajectories of offending for sexual offenders are distinct from general offending trajectories, and whether different trajectories for sexual offending exist. Although not employing a group based model, Hanson (2002) analysed trajectories of recidivism risk, and found that the aggregated age-crime curve for sexual offenders was bimodal, with a peak before the age of 20 corresponding to the peak of offending in the general population, and a second peak in the mid- to late-30s. He also found differences between rapists and child molesters in the rate of decline of recidivism risk. He found that the risk of recidivism for child molesters did not begin to decline until well into middle age. Lussier (2010) fitted a group based trajectory model to a sample of sexual offenders, and found groups of "late bloomers" whose offending appeared to accelerate into middle age. He also found that child molesters were more associated with this late onset group.

Apart from post-hoc analyses of classes to explore their constitution, studies of criminal career trajectories have not tended to address specialisation/versatility, and vice versa. The specialisation/versatility debate and the criminal career trajectories debate cannot be artificially separated in this way. If, as both Harris et al. (2009) and Hanson (2002) claim, there are important differences between rapists and child molesters, both in terms of specialisation and trajectories, then the interaction between frequency and specialisation needs to be understood. Furthermore, if, as both Soothill et al. (2000) and Harris et al. (2009) claim, versatility is the rule, then perhaps the classification of offenders as child molesters or rapists, which is the starting point for comparing their trajectories, is dubious to begin with.

The bimodal age-crime trajectory noticed by Hanson (2002) is another empirical observation that requires a blend of approaches to understand. Is this bimodal age crime curve caused by the aggregation of different offenders with different single peaked trajectories? Or are there individual offenders who follow double-peaked offending trajectories? In either of these cases, are the different peaks associated with apexes of the occurrence of different types of crime? If yes, then perhaps the appearance of multiple trajectories is caused by the aggregation of different types of crime, and would disappear if the crimes were disaggregated?

1.b Research questions

It is from the intersection of the specialisation/versatility debate and the field of sexual offender trajectory research that the substantive research questions for the present study arise. These can be summarised as follows:

1. Bimodality of trajectories:
    - Is the observed bimodality of trajectories for sexual offenders unique to sexual offenders? and if so:
    - Is it within- or between- offender?
    - Is it within- or between- crime type?
    - i.e. Can it be explained by disaggregating offenders into classes, or by disaggregating criminal activity into crime types?
2. Association between child molestation and rape:
    - Is there a strong separation between rapists and child molesters?
    - Do offenders commit only one of these over the life course, or both?
    - Do offenders specialise in one at one point in their lives, and another at another point?
3. Association of non-contact sexual offending with serious offending:
    - Are rapists, or child molesters more likely to commit non-contact sexual offences?
    - Does non-contact sexual offending tend to precede serious sexual offending, in a pattern of escalation, or is there no evidence of this?

The first set of questions relate to the bimodality of trajectories. Answering these questions will require the use of specifications for the group based trajectory model that allow bimodal trajectories to be fitted, since the currently common quadratic and cubic polynomial trajectories would hide bimodal trajectories if they exist.

The second and third sets of questions relate to longitudinally dynamic aspects of the criminal profiles (mix of crimes) of offenders, and require multivariate methods to answer. The suggestion that specialisation might be local, within an overarching pattern of versatility over the life course, demands models that can distinguish between within-period association of the occurrence of different crimes, and within-individual association.

1.c Aims of the study

The purpose of this study can be separated into three broad aims:

1. To address the substantive research questions:
    (a) Bimodality of trajectories;
    (b) Association between child molestation and rape;

(c) Association of non-contact sexual offending with serious offending.

2. To implement and evaluate four approaches to jointly analysing criminal career trajectories and profiles (mix of crimes):

    (a) Optimal matching and non-probabilistic clustering;

    (b) Constrained and unconstrained multivariate trajectory models;

    (c) A Poisson-log normal factor trajectory analysis.

3. To investigate, and if possible account for, inferential problems caused by the circumstances of construction of the sample.

The second and third aims arise necessarily from the thorough pursuit of answers to the substantive research questions. The association of sexual offenders, and child molesters in particular, with late-onset criminal career trajectories, as suggested by Lussier (2010), will be assessed using optimal matching and multivariate group based models. In addition, the study will assess the existence of multi-modal trajectories, as predicted by Hanson (2002).

The Poisson-log normal factor trajectory model, and a posterior log-linear analysis of class membership probabilities, will be used to investigate the extent to which certain categories of sex offender are separate, or the extent of overlap between types, and will allow for the possibility of offenders following multiple trajectories for different crime types.

For inferences from any analyses to be valid, the sample must be an unbiased sample of a population of interest, and we will also investigate whether this is the case, whether adjustments can be made to models to account for any biases, and the extent to which substantive findings can be generalised.

1.d Structure of the dissertation

The next section will introduce the sample, describe the background to the collection of the data, and present some general summary measures of the properties and composition of the dataset. The main motivations for the use of the sample will be explained, as well as some limitations of the sample that must be accounted for if it is to be appropriate to draw inferences from it.

The rest of the study will be structured as three independent analyses: in the first analysis, non-probabilistic data mining techniques will be applied to classify and visualise the criminal careers as sequences of events. This analysis will serve as a more thorough exploratory data analysis, and aid the generation of hypotheses for the following analyses.

The second analysis will consist of fitting two group based trajectory models. The first will involve fitting a group based trajectory model to frequencies of total offending, as a benchmark against which to compare the multivariate analyses. In this section the specification of the group based trajectory model will be laid out, as well as some general methodological decisions that apply to all of the GBTMs fit in this study. In the second part of the second section, frequency of offending will be disaggregated into two sub-frequencies - sexual offending and general offending, and GBTMs will be fitted to both frequencies. Two types of multivariate GBTM will be fitted to these dual frequencies and compared.

In the third analysis, a factor model for count data will be introduced, as a way of reducing the dimensionality of the crime frequencies without making *a priori* decisions about aggregation, and preserving the covariance structure of the crime categories. It will be shown that the factor model can be used to produce statistics called factor scores, and these will be used in trajectory models to determine whether trajectories have simple structure once crime types are accounted for.

In the final section, we will draw upon the results of the three analyses to determine if the aims of the study have been met, and if answers to the questions posed in section 1.b can be answered. The limitations of the study, and suggestions for future studies to build upon this one, will also be discussed.

2 The Sample

The sample is a set of 824 male offenders who were referred to the Massachusetts Treatment Center (MTC) for Sexually Dangerous Persons, part of the Bridgewater State Hospital in Massachusetts, USA, between 1959 and 1984. All of the men had committed at least one serious sexual offence involving contact with a victim (rape, child molestation, contact sexual offences). The offenders were referred to the MTC via the legal system under laws for "sexual psychopaths". In many cases the referral was made at the time of conviction for the index offence(s), which would have been serious sexual offences, but in some cases the referral was made a long time after the index offences when the subject had been brought before the courts for another reason. Approximately half of the men were admitted as inmates into the treatment center; the rest were judged not to be sexually dangerous and were either released, or sent back into the prison system to serve their sentences.

This is therefore a convenience sample, and the question of its generalisability to the subpopulation of serious sexual offenders is not taken for granted, but is not addressed explicitly by this project. Despite the limitations caused by the circumstances of its creation, the dataset is nevertheless an incredibly rich source of criminal histories for sexual offenders. Each offender has an average of 15 offences, and the largest number of offences for a single offender is 96.

For each offender the date of birth has been recorded, as well as an indicator for whether he was admitted or released, and various other pieces of personal information that are not employed in the study. The dataset lists all recorded crimes for each offender, from the age of seven to the time of referral to the MTC. There is also information on offences committed after release from the treatment center. This follow-up data was not used because the two sets of offending data are separated by an indeterminate period of incarceration in the treatment center, which could bias the estimated trajectories.

Offending histories consist of a date for each offence, which was usually the date of conviction or the date of charge, but in some cases exact dates were not available, and only the month or year of the offence was available. Because in most cases the dates are not the actual dates of offence, there is some clustering in offence dates.

The dataset was used in two forms during the three analyses. For the optimal matching the offences were not aggregated into time periods, and were used in long person-event form. For the group based and factor models offences were aggregated into counts by five-year period for each individual, with a count for each category of offence, sub-totals for sexual and general (non-sexual) offending, and a total for all offending in each period.

The type of offence is recorded as a nominal variable with, originally, 20 categories. These 20 categories were mapped to eight categories for the optimal matching and factor analysis, and two broad categories - sexual offences and general (non-sexual) offences - for the dual trajectory analysis. Table 1 shows the frequency of offences by the original categorisation, and the mappings to the eight categories. Underlined categories are included in the sexual offending total, whereas the rest are included in the general offending total. One category, homicide, was not used in the factor analysis, because there were too few occurrences. Of the explicitly sexual offences, all were mapped directly, apart from "contact SO" which was merged with rape because of the small number of occurrences.

It can be seen in Table 1 that by far the greatest number of offences were child molestation, which exceeds the combined total of the other sexual offences. This is perhaps related to the age of the sample; in the mid-twentieth century some categories of rape, such as date rape or marital rape, were considered less serious than they are today and were prosecuted rarely. In contrast, child molestation has always been considered a serious offence, and a serious form of sexual deviancy. The second most numerous type of offending is motoring offences. Rather than removing these and other possibly trivial offences from the dataset, they were retained so that the relationships between serious sexual offending and low-level law-breaking could be examined.

The third piece of information for each offence was the disposition (sentence) for the offence. These were originally categorised into twenty categories. This information was only used in the study in the form of an indicator of incarceration in the current period. Lag variables were also created indicating incarceration in each of the previous three periods for each offender. These were used in the zero-inflation model to account for an increased risk of zero inflation in periods following a custodial sentence. An exposure indicator was calculated, equal to five in periods with no incarceration, and reduced proportional to the time that the offender was at liberty during other periods.

Table 1: Frequencies of the twenty original offence types, mapped to the seven aggregated offence types.

Offence Type	Mapped to	Count	Offence Type	Mapped to	Count
Child molest.	Child molest.	2679	Misc.	Other	300
Motoring	Other	1864	Abduction	Other	283
Assault	Assault	1226	Property	Property	272
Breaking & ent.	Breaking & ent.	997	Justice/ milit.	Other	233
Rape	Rape	945	White-collar	Other	207
Theft	Property	912	Dangerous act.	Other	204
Non-contact SO	Non-contact SO	805	Contact SO	Rape	178
Alcohol	Other	690	Weapons	Other	120
Public order	Other	659	Drugs	Other	87
Robbery	Property	309	Homicide	Homicide	79

2.a Exploratory analysis

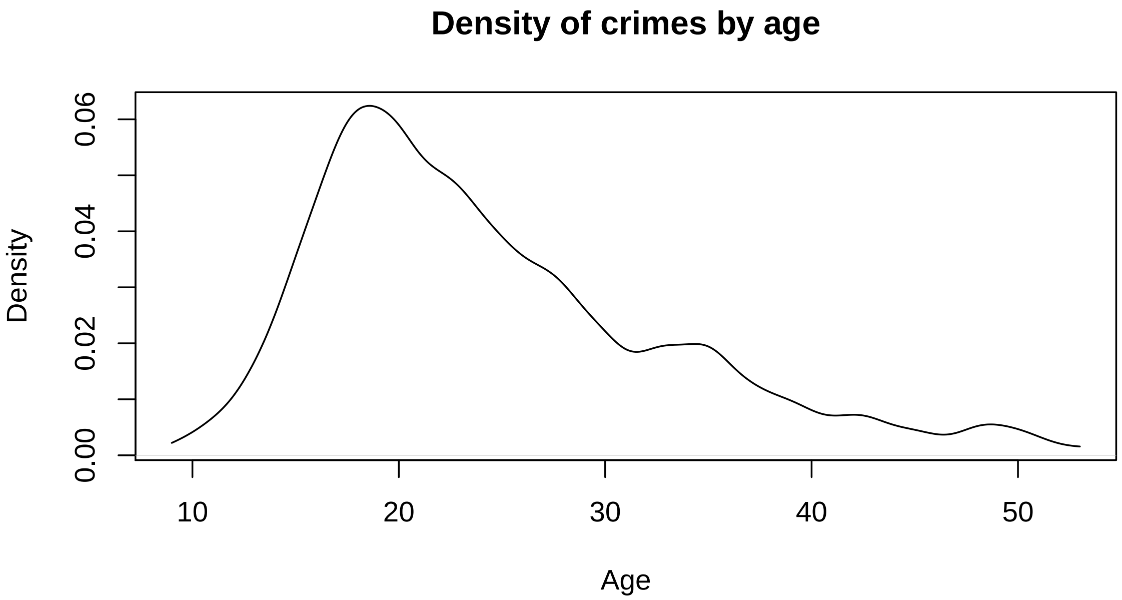


Figure 1: *Density plot of the number of crimes in the unaggregated dataset, by age. The kernel is gaussian and the bandwidth=1.*

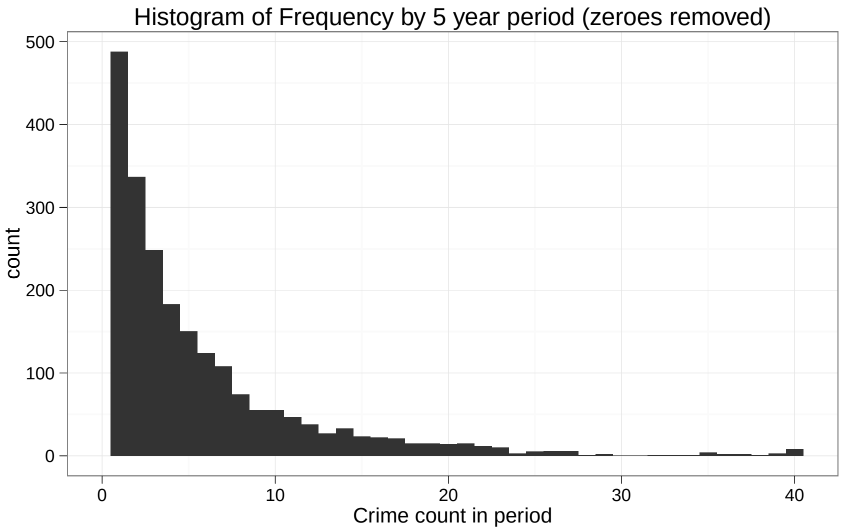


Figure 2: *Histogram of frequency of crimes in each five-year period. 1653 pre-referral zeroes have been removed.*

Some basic exploratory analysis was conducted on the dataset. Firstly, missing data was checked. There is minimal missing data on the offences themselves. There are less than

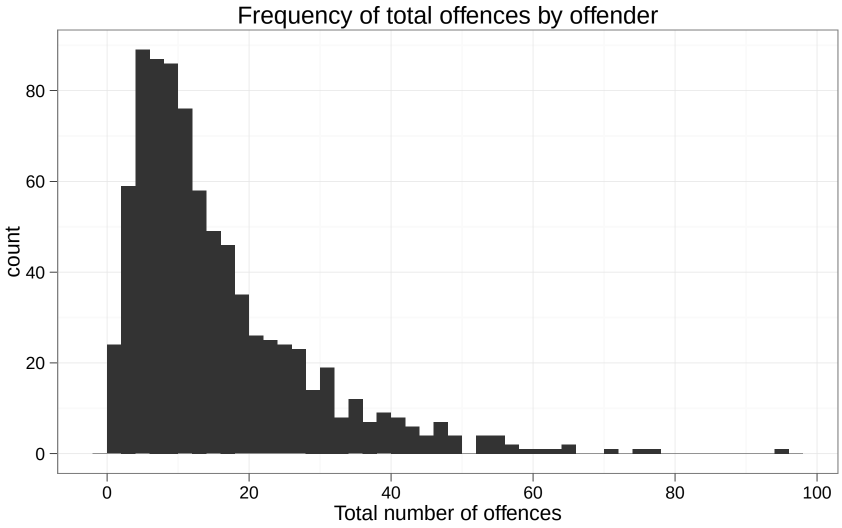


Figure 3: *Histogram of total offences committed by offenders.*

2% missing dates (either date of birth or date of offence) and no missing offence types. The offences with missing dates were deleted, after checking that they did not appear to be associated with any particular types of crime. In total, there is complete data for 13,049 offences.

Secondly, the overall distribution of counts was plotted at the level of each offender, and also at the level of five-year periods. It was found that offenders have an average of 15.1 offences throughout the observation period (see Figure 3). The average number of offences in each five-year period before referral is 3.3 (see Figure 2). Figure 1 shows the empirical density of offences by age, plotted using gaussian kernel density estimation. The peak is before the age of 20, as in the archetypal age-crime curve although there is a second smaller peak in the mid-30s at around the age where a second peak would be predicted by Hanson (2002).

2.b Non-ignorable and unobserved missingness

There are two forms of missingness that are problematic in the sample. The first is a form of unobserved missingness due to unrecorded periods of incarceration. The second is a form of non-ignorable missingness due to right censoring.

The problem of unobserved missingness due to incarceration has been considered previously with regards to GBTMs. Eggleston et al. (2004) underlined the importance of treating periods of incarceration as missing data, to avoid downward bias in estimates of crime frequency due to periods of incarceration being treated as periods of no offending. The Massachusetts dataset, in common with many other criminal careers datasets, does not contain information on periods of incarceration; in effect the missingness indicator is itself missing. A common approach to accounting for so-called "structural zeroes" in GBTMs has been to employ a zero-inflated Poisson (ZIP) model (Nagin and Land, 1993)(see Section 4.b.iii).

The second missingness problem is related to the sampling process. In the Massachusetts dataset, the follow up period ends when the offender is admitted to the MTC. Because of this, all observations are right censored at the point of referral. By construction, there are no participants who die or desist from criminal activity before the end of the follow up period, and referral is usually at or near a period of frequent offending. This means that the sample is unsuitable for modeling desistance from crime, because all criminal careers are truncated before desistance.

The problem is exacerbated by a steep decline in the size of the active sample from around the mid-20s onwards (see Table 11a in Section 4.b.iii). This means that not only are those offenders who remain bound to be active offenders, but also the frequency of their offending will have a disproportionate effect on the height of the trajectories they contribute to. The use of group-based models means that it is difficult to calculate exactly how many offenders contribute to each trajectory, because contributions are weighted by posterior class membership probabilities.

The problem of non-ignorable dropout in GBTMs has been addressed by Haviland et al. (2011), who outlined a generalisation of the model to incorporate dependence of the dropout process upon class membership and other covariates. This generalisation is designed to alleviate bias in the estimation of class membership probabilities. However, this extension would not address the unsuitability of the sample for modeling desistance or the inevitably biased estimates of trajectories towards the end of the period, which are a consequence of the sample's construction. The latter problem has been addressed to some extent by truncating the follow up period at an age where the active sample size is still reasonable (the period 49-53).

In Section 4.b.iii the zero-inflated model for periods of unobserved missingness is outlined. It is explained that extending this model not only to intermittent missingness, but also missingness after referral, has been used to counteract bias in the estimated trajectories.

3 Analysis One: Optimal Matching

3.a Motivation

Sequence mining methods were incorporated into the study for two reasons. Firstly, since they require no data aggregation and almost no transformation, the complexity of the data was not hidden. Secondly, optimal matching techniques are non-statistical, and as such rely on very different set of assumptions to group based trajectory models; it is therefore useful to compare the results of applying these two techniques as a way to understand the implications of the assumptions being made in both.

The analysis was conducted using the R package[1], TraMineR (Gabadinho et al., 2011) which incorporates tools for the analysis of state and event sequences. Although primarily oriented towards state sequences, there are nevertheless specific tools for use with event sequences, and Studer et al. (2010) has recently extended the capabilities of TraMineR with the addition of a function to calculate a dissimilarity matrix for event sequences, based on the concept of "edit distance" from one sequence to another. Once a dissimilarity matrix has been calculated, widely recognised clustering algorithms, such as agglomerative nesting and partitioning around medoids, can be used to cluster the sequences.

The application of methods of optimal matching in the social sciences have been criticised because it is not clear how the conceptual model of edit distance translates well from text and genetic research to the domain of criminal careers (Levine, 2000). Additionally, they require the researcher to make subjective decisions about costs of various edit operations. Also, McVicar and Anyadike-Danes (2010) asked whether there was an application of optimal matching in the social sciences that could not be better achieved with other methods. It will be demonstrated that OM methods are useful for exploratory analysis and should be regarded as a complement to model-based analyses, although we will not address the question of the arbitrariness of the parameters.

3.b Methods used in the optimal matching analysis

3.b.i Optimal matching and edit distance

In event sequence analysis, a criminal career is represented as a string of ordered events, together with the time separating each pair of consecutive events. An example might be the following sequence:

$$\rightarrow^{17} (\texttt{Assault}) \rightarrow^{2} (\texttt{Theft}) \rightarrow^{1} (\texttt{Rape}) \rightarrow^{5}$$

Where the numbers above the arrows are the length of the period separating two events, in years (Studer et al., 2010). It can be seen that the sequence is augmented with the length of the period before the first event (age of onset) and the length of the period in between the last event and the end of the observation period. It is therefore possible to define variable observation periods and right censoring.

Optimal matching is based upon the premise that any sequence of events can be mapped to any other sequence of events by a series of transformations, or "edits" of the sequence. Different variations on the method allow for different types of operation. The method used

[1] R Core Team (2012).

in this study allows for three types of operation: insertion of an event; deletion of an event; and translation of an event by a unit of time (either forwards or backwards). Each type of operation is associated with a cost, φ, which can depend upon the type of event, in the case of insertion/deletion costs. These costs are chosen by the researcher, and it has been suggested that the cost of an insertion/deletion should be inversely proportional to the probability of occurrence of the event in the sample (Studer et al., 2010).

The unscaled edit distance or Levenshtein distance $d(A, B)$ between two sequences of events A and B is equal to the minimum total cost of transforming one sequence to be equal to the other. The cost of insertion of an event is constrained to be equal to the cost of deletion of the same event, otherwise the distances will not be symmetrical (i.e. $d(A, B) \neq d(B, A)$).

The unscaled edit distance is a euclidean distance, in that it respects the triangle inequality. However, the distance depends upon the length of the sequences, with two short sequences likely to be closer to each other than two longer sequences. For this reason, the edit distance is scaled so that each distance is proportional to the maximum possible distance between the two sequences. The scaled distance lies in the range 0-1, where 1 is the maximum possible distance, and is also a euclidean distance.

Edit distances are calculated using a dynamic programming algorithm, the output of which is an $n \times n$ dissimilarity matrix containing the scaled edit distances between each pair of sequences (criminal careers). The dissimilarity matrix can then be used as a summary measure in a subsequent data analysis. The most common use of dissimilarity matrices calculated in this way is to subject them to some form of clustering, to uncover hidden structure in the event sequences.

3.b.ii Non-probabilistic clustering techniques

Non-probabilistic clustering techniques are used to find categorical structure in data that has not been pre-categorised. This task is achieved using many different methods, two of the most popular of which are heirrarchical clustering, and partition-based clustering.

Heirrarchical clustering begins with the unclustered data, and proceeds by clustering observations or clusters of observations that are close together, until all of the data has been encompassed in one group. At each cycle, the algorithm only makes one join, that of the two observations or clusters that are closest in that cycle. The closeness of two clusters can be measured either from the centres of the two clusters, or from the nearest two observations in the two clusters, or from the farthest two observations in the two clusters. A popular method, Ward's method, uses the increase in variance of each cluster as a measure of distance at each stage. The result of heirrarchical clustering is not one set of clusters, but a tree-like structure of successively aggregated clusters. The researcher can examine the heirrarchical structure using a visualisation called a dendrogram, and decide upon the most desirable clustering by eye.

In partition-based clustering methods, in contrast, the researcher must decide upon the number of clusters, k, beforehand. A popular partition-based method is called Partitioning Around Medoids (PAM). In PAM, a set of k representative observations are chosen, called medoids, and other observations are clustered according to their closeness to each of these medoids. The goodness of the clustering produced by a given set of medoids is measured by the sum of the dissimilarities of all objects to their nearest medoid. The algorithm proceeds by swapping one medoid at a time until the sum of dissimilarities cannot be reduced further. The output of PAM is a cluster assignment for each observation into one of the k clusters, and a set of k representative medoids.

A summary measure of the goodness of a k-cluster solution is the average silhouette width of observations, that can be used to choose between different values of k. The silhouette width of an observation, y_1 is defined as follows. Firstly, the average distance between y_1 and all other y_j in the same cluster, c_1, is calculated:

$$D(y_1, c_1) = \frac{1}{n_1} \sum_{j \in c_1} d(y_1, y_j)$$

Then the average distance between y_1 and all y_j in the nearest neighbouring cluster:

$$D(y_1, c_2) = \min_{i \in k}(D(y_1, c_i)) = \min_{i \in k}\left(\frac{1}{n_i} \sum_{j \in c_i} d(y_1, y_j)\right)$$

The silhouette width of y_1 is then:

$$s(y_1) = \frac{D(y_1, c_2) - D(y_1, c_1)}{\max\left(D(y_1, c_1), D(y_1, c_2)\right)}$$

Different clustering algorithms were compared. The heirrarchical clustering tended to produce a combination of a few very large clusters and some single-observation clusters, which was undesirable for the purposes of the study. The TraMineR package contains a function that calculates a PAM solution, using the centroids of the k level of a Ward heirrarchical clustering as start values. This function was compared to the default PAM function and produced the best average silhouette widths, so this function was used to calculate the clusters.

3.b.iii Method

As stated previously, very little data manipulation was required prior to calculating optimal matching distances. The crimes were unaggregated, and age of the offender at the time of offence was used as the time variable. The end time for each sequence was set as the time of the last offence.

Following Studer et al. (2010) the insertion/deletion costs were chosen to be inversely proportional to the proportion of occurrence of each type of crime in the sample. This has the effect of balancing the influence that each type of crime has on the distance, and hence the clustering, otherwise clusterings would be dominated by small differences in the distribution of the most common crimes. Sexual offences were weighted by a factor of three to reflect their relative importance in the analysis. The resulting costs were scaled so that the smallest cost, that for "Other" offences, has a cost equal to one. Table 2 shows the costs used. The translation cost was set equal to 0.3, so that a translation of six-seven years would be equivalent to a deletion and an insertion of a completely new event in the case of the most common category of crime.

Table 2: Insertion/Deletion costs for the different crime categories.

homicide	assault	rape	child	noncont	other	property	breakent
58.82	3.79	12.42	5.19	17.31	1.00	3.11	4.66

3.c Results of the optimal matching analysis

Table 3 shows the average silhouette widths for two to six clusters from the PAM clustering with Ward start values. The maximum value is for the four-cluster solution, which is therefore selected. The average silhouette widths for each cluster are given in Table 4, which shows that there are no clusters with very poor separation, and one small cluster with very good separation from the others.

Table 3: Choice of PAM/Ward solution by average silhouette width

No. of clusters	2	3	4	5	6
Avg. Silhouette width	0.11	0.12	0.13	0.12	0.12

Table 4 also shows the average criminal career characteristics of the four clusters. Although all of the clusters have different averages, in general there seems to be evidence of two “super clusters” grouping clusters one and two, and clusters three and four. The first two clusters have later ages of onset, later ages of referral, and a lower intensity of criminal activity (crimes/period length). Length of career and total crimes do not exhibit this two-way structure, however.

Table 4: Properties of the four PAM/Ward clusters.

		Average:					
Cluster	%	Silhouette width	Age of onset	Age of referral	Length of career	Total crimes	Crime Intensity
1	24.4	0.137	20.1	31.7	11.6	16.3	1.4
2	31.4	0.098	21.5	32.1	10.6	11.7	1.1
3	36.5	0.128	17.2	25.7	8.5	16.5	1.9
4	7.6	0.281	17.2	29.2	12.0	19.5	1.6

The criminal profiles of the four clusters are summarised in Table 5. It can be seen that the four clusters are similar in their commission of “Other” crimes, which comprise around one third of crimes for all clusters. Apart from this, the criminal profiles appear to follow the same super clustering as the criminal career statistics. Clusters one and two are characterised by the commission of child molestation, whereas clusters three and four commit more assault, rape, property, and breaking and entering offences. There are a couple of crimes which are dominated by one cluster: cluster one commits 86% of all non-contact sexual offences, and cluster four commits all of the homicide offences. The separation of the murderers was a consequence of the large insertion/deletion cost given to homicide, but was not inevitable, and would not have occurred if murderers were not otherwise quite homogenous in their offending sequences. Although the murderers have a very similar profile to cluster three on the other crimes, it is notable that they commit less rape per person than members of cluster three, and more child molestation.

Table 5: Number and proportion of crimes committed by the four PAM/Ward clusters.

	Homicide	Assault	Rape	Child	Non-cont.	Other	Property	B & E	Total
Cluster One: 24.4%.									
No.	0	241	113	934	698	1069	259	192	3506
%	0	6.9	3.2	26.6	19.9	30.5	7.4	5.5	100
Cluster Two: 31.4%.									
No.	0	179	18	1435	23	1218	232	143	3248
%	0	5.5	0.6	44.2	0.7	37.5	7.1	4.4	100
Cluster Three: 36.5%.									
No.	0	626	880	214	63	1902	809	538	5032
%	0	12.4	17.5	4.3	1.3	37.8	16.1	10.7	100
Cluster Four: 7.6%.									
No.	79	180	112	96	21	458	193	124	1263
%	6.3	14.3	8.9	7.6	1.7	36.3	15.3	9.8	100

To visually summarise the criminal careers of the four clusters, both empirical trajectories (mean number of crimes per year) and Kaplan-Meier curves have been plotted in Figures 4 and 5. The two types of plot are complementary: the empirical trajectories can be interpreted in a way similar to trajectories from a group based model; and the Kaplan-Meier curves represent the probability of members of the cluster not having committed at least one crime of each type at each point. The Kaplan-Meier curves are therefore useful for analysing age of onset, and to some extent shed light on specialisation, since they show to what extent offenders are exclusive in their offending. The tick marks on the survival curves indicate when an observation is censored without having committed that crime.

The two-way super clustering is evident in Figure 4. The first two clusters are characterised by the commission of child molestation over a long period, whereas the second two clusters are more associated with assault, breaking and entering, and rape, with a shorter and earlier period of high activity. The survival curves in Figure 5 show, however, that of those whose criminal careers last to the end of the period, most of the offenders in all clusters will have committed child molestation at least once.

The plotted trajectories also make clear how the clusters are different within each super cluster; cluster two is associated almost exlusively with child molestation, whereas cluster one is active in assault, rape, breaking and entering, and particularly non-contact sexual offending. Cluster three, the largest cluster, is most associated with rape and assault, whereas cluster four is also associated with child molestation and murder. Cluster two is the most specialised, whereas cluster four, the murderers, are the least specialised and have a reasonable probability (at least 0.4) of having committed all crimes except non-contact sexual offences by the age of 30. By the age of 25, around half of this cluster have committed murder.

The survival curves suggest that in the second two clusters, those who are offending in adolescence usually begin with breaking and entering, and progress to assault and then rape. In the first two clusters there is not much evidence of progression, and members of these clusters seem likely to begin offending by committing child molestation.

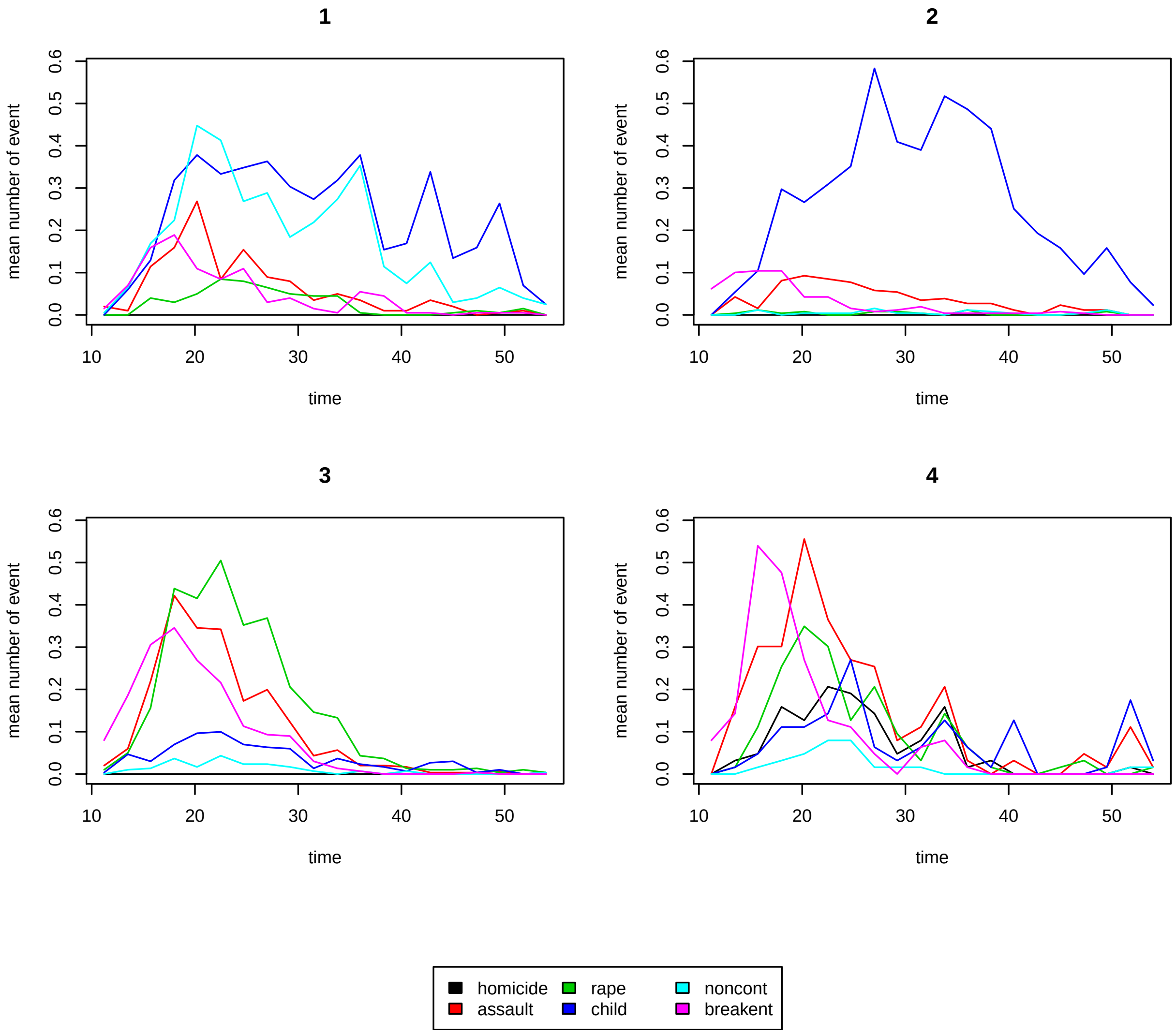


Figure 4: *Mean number of events per year for each crime type for the four PAM/Ward clusters. "Other" and "Property" lines are not plotted.*

3.c.i Model checking

In order to provide some kind of sense-check of the clusters produced by the PAM/Ward algorithm it was necessary to somehow visualise the space over which the optimal match-

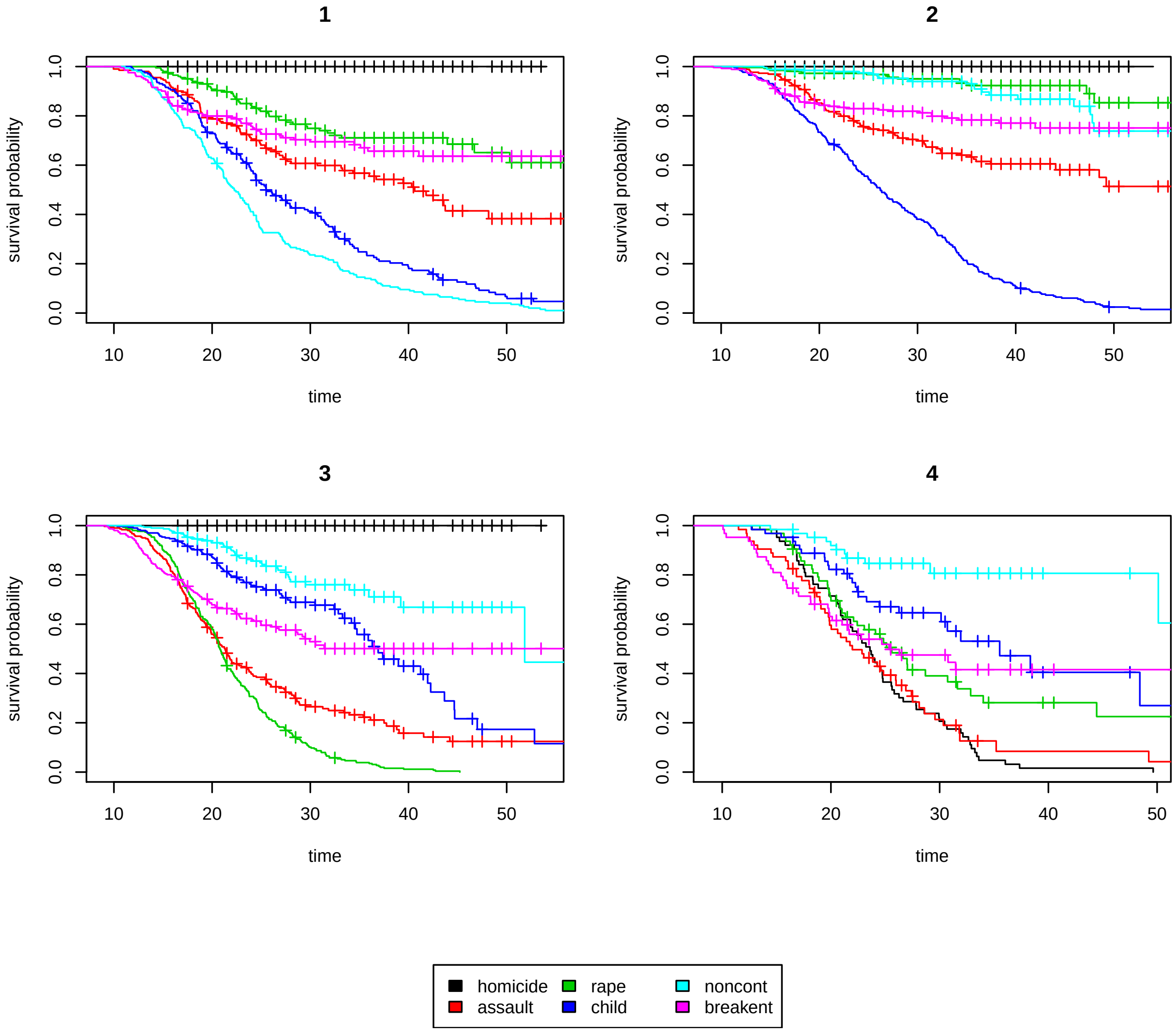


Figure 5: *Kaplan-Meier survival curves for the first occurrence of each crime type for the four PAM/Ward clusters. "Other" and "Property" lines are not plotted.*

ing produced distances, and over which the sequences were clustered. The distance matrix produced by the optimal matching was mapped to a two-dimensional space using multidimensional scaling, and the clusters were plotted in Figure 6. There appears to be quite a well-defined cluster structure to the sequences, with at least three clusters visually discernable in the plot. Although largely consistent, the clustering produced by the PAM/Ward algorithm does not correspond exactly with what would be expected from the plot, suggesting that the topographical mapping produced by multi-dimensional scaling does not correspond exactly to that produced by the optimal matching. The multidimensional scaling is constrained to two dimensions which might explain this.

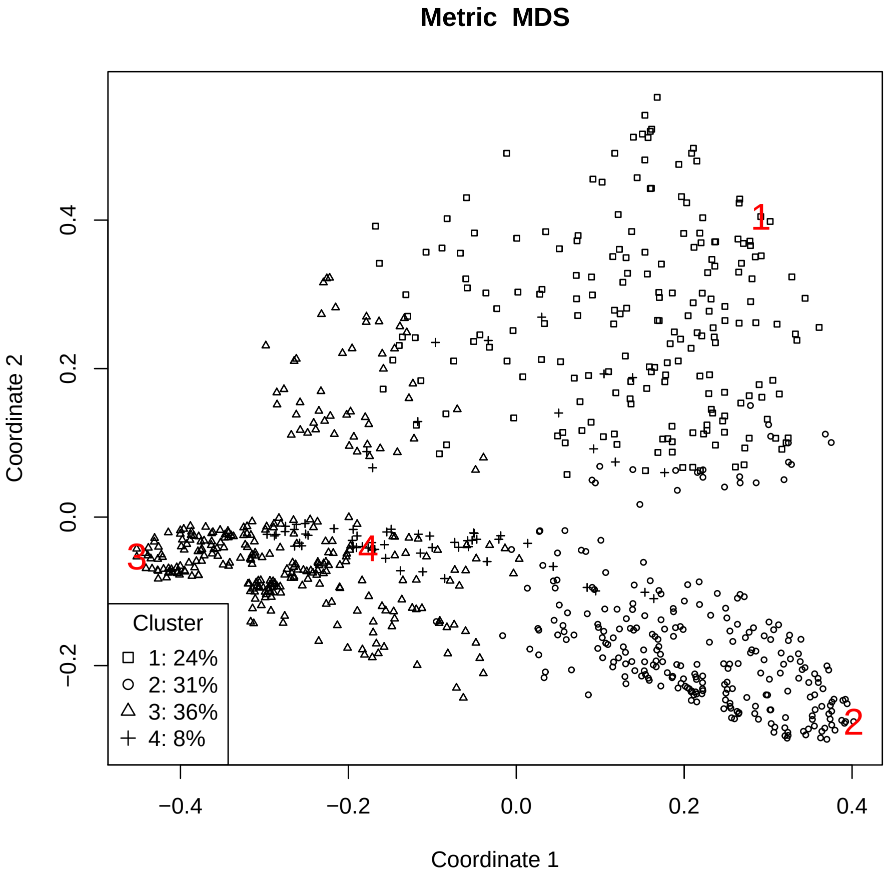


Figure 6: *Sequences mapped to 2D space using multi-dimensional scaling. Symbols represent PAM/Ward clusters. Large red numbers show positions of the four medoids*

3.d Evaluation of the optimal matching results

The combination of optimal matching and algorithmic clustering has provided a rich summary of both trajectories and profiles of offenders in the sample. The four class solution from the PAM/Ward clustering indicated at the very least a strong two-way structure, separating rapists from child molesters. The super cluster of child molesters, comprising 56% the sample, committed 88% of all child molestation, and the super cluster of rapists, comprising 44% of the sample, committed 72% of all rape. However, the Kaplan-Meier curves suggest that child molestation was probable for all sample members whose criminal careers lasted long enough. In general, offenders appear to be versatile as long as their criminal careers lasted long enough, with most having a greater than 50% probability of having committed more than one type of offence by the age of 30, apart from those in cluster two.

The discovery of a cluster of murderers was a consequence of the insertion/deletion cost assigned to murderers, which reveals a limitation of the method - that it is dependent upon the choice of optimal matching costs. This is not a limitation, however, if optimal matching is viewed as a way of summarising complex data rather than as a way of making inferences on phenomena outside the dataset. The separation of all those committing homicide was valuable because it revealed that on the whole those committing murder are generalists, and are more associated with rape than child molestation. The value of optimal matching is that it summarises data using rules determined by the researcher, but providing results not necessarily foreseen by the researcher, and is therefore a good impetus for hypothesis generation.

Varying observation periods were not taken account of by the empirical trajectories, but were taken account of in the calculation of the dissimilarity matrix, and therefore in the clustering. The Kaplan-Meier curves take account of right censoring by definition, as they show the proportion of those in the at-risk group who have not yet committed the crime, and the tick marks indicate when an observation is censored without having committed that crime.

4 Analysis Two: Group Based Trajectory Models

4.a Motivation

However useful they may be for data exploration, non-probabilistic methods do not allow the extension of inferences on a phenomenon outside the data. In order to shed light on the research questions, it is necessary to model the processes that generated the data, and model the uncertainty in those processes. Finite mixture models, and group based trajectory models in particular, provide a flexible semi-parametric framework to model complex phenomena whilst keeping assumptions to a minimum. We will introduce multivariate extensions to the group based trajectory model with the aim of imitating the ability of optimal matching to jointly model trajectories and profiles, in a statistical framework.

4.b Group based trajectory models

Group-Based Trajectory Models are longitudinal regression models for counts[2]. Typically, counts are modeled using Poisson or negative binomial distributions, or their zero-inflated (ZI) variants. The "trajectory" part of their name comes from the use of linear and polynomial terms for time (usually biological age). The "group-based" part of their name comes from the fact that GBTMs are a type of finite mixture model.

Finite mixture models are a popular class of models that approximate an underlying distribution of a phenomenon of interest by a combination of distinct distributions. Finite mixture models have two complementary interpretations and justifications: the first is as a convenient semi-parametric way of fitting complex distributions that are not easily fitted by known distributions; and the second is as a way of inferring the existence and membership of unmeasured subpopulations in the sample.

In the "semi-parametric fitting" interpretation, a random vector of variables for the j'th member of the sample, $\mathbf{Y}_j$ $(j = 1, \ldots, N)$ is thought to be distributed according to an unknown complex distribution, $f(\mathbf{y}_j)$[3] which might be skewed, overdispersed, or multimodal.

[2] Usually, although a linear version for factor scores was also employed in this study

[3] Which may be either a continuous density or discrete probability mass function

This unknown distribution can be approximated by:

$$f(\mathbf{y}_j) = \sum_{i=1}^{k} \pi_i f_i(\mathbf{y}_j)$$

where π_i are mixing proportions or weights summing to one, and $f_i(\mathbf{y}_j)$ are the component densities $(i = 1, \ldots, k)$, the individual parametric distributions of which the mixture is comprised. Using this formulation, almost arbitrarily complex distributions can be fitted, depending on the choice of the component densities and mixing weights. The component densities do not need to come from the same family.

In the "latent class" interpretation, the random sample, $\mathbf{Y}$ is assumed to be sampled from several distinct subpopulations, each with its own set of unknown parameters for the distribution(s) of the variables of interest. The subpopulation membership of the j'th member of the sample is denoted by $\mathbf{Z}_j$, a vector of component membership or classification indicators for each of the components i $(i = 1, \ldots, k)$. Assuming the latent class model is valid, each member of the sample can only belong to one sub-population or component, so only one of the z_{ij} can be one, and the rest are zero. However, the values of $\mathbf{Z}_j$ are not known.

This formulation allows mixture models to be fitted as "missing data" problems, using estimation procedures such as Expectation-Maximisation (EM) or Data Augmentation MCMC. In this interpretation also, the π_i take on another meaning as the proportions of the sample who belong to each of the k latent classes, and consequently as the prior probabilities of class membership for each of the participants. This interpretation allows the posterior probabilities of class membership conditional upon the observed data $\tau_{ij} = \Pr(z_{ij} = 1|\mathbf{y}_j)$ to be calculated for each individual. These can be used to assign members of the sample to components, usually on the basis of the modal value of $\boldsymbol{\tau_j}$. These modal assignments can be denoted $\hat{\mathbf{Z}}_j$.

These interpretations are not entirely distinct, and in many applications their use is justified from both perspectives; GBTMs are a good example of this. Indeed Nagin (2005) originally justified the GBTM approach as a semi-parametric way to overcome the over-dispersion usually present in frequencies of criminal offending[4]. However, the continuing focus in the literature on plotting and labelling the component trajectories, and the strong links to the typological work of theoretical criminologists such as Moffitt suggests that the main reason for the continuing popularity of GBTMs is their power to classify complex longitudinal datasets.

In GBTMs, the general finite mixture model is extended to allow for repeated observations. This is achieved by only letting $\boldsymbol{\tau_j}$ vary between individuals and not within individuals and imposing an assumption that observations are independent of each other conditional upon component membership. The general form of the GBTM is given by:

$$p(\mathbf{Y}_j) = \sum_{i=1}^{k} \pi_i p_i(\mathbf{Y}_j|\boldsymbol{\theta}_i)$$

where each $p_i(\mathbf{Y}_j|\boldsymbol{\theta}_i)$ is a count distribution with parameters $\boldsymbol{\theta}_i$. Due to the assumption of local independence of observations, $p_i(\mathbf{Y}_j|\boldsymbol{\theta}_i) = \prod_{t=1}^{T} p_i(y_{jt}|\boldsymbol{\theta}_i)$. If the count distributions are

[4] See Section 4.b.ii

Poisson, then:

$$p_i(y_{jt}|\texttt{Age}_{jt}, \mathbf{z}_j, \boldsymbol{\theta}_i) = \frac{\lambda_{it}^{y_{jt}} e^{-\lambda_{it}}}{y_{jt}!}$$

where $\texttt{Age}_{jt}$ denotes any terms for age included in the design matrix, and $\lambda_{it} = E[y_{jt}]$ is the rate parameter for component i at time t. The relation of λ_{it} to age covariates and parameters is specified using a log-link so that:

$$\log(\lambda_{it}) = \mathbf{x}_{jt}^T \boldsymbol{\beta}_i$$

Note that in this case $\boldsymbol{\theta}_i = \boldsymbol{\beta}_i$.

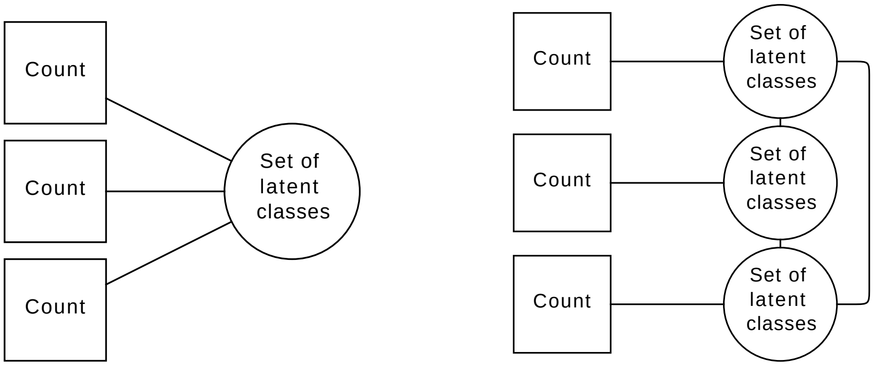
4.b.i Modeling multivariate data

GBTMs can be extended to situations where there are several variables, each measured repeatedly for each individual. In the context of this study, the variables are counts of different categories of crime. GBTMs can be extended to handle multivariate random vectors in three ways. The first way is by modelling the components as multivariate distributions; this is done in the case of model-based clustering, where multivariate normal components are used. However, this is problematic for counts, because multivariate models for counts are limited and difficult to apply.

The second method exploits the fact that if variates are assumed to be independent, the joint density for each component reduces to a product of univariate densities. Using this method, variables are assumed to be uncorrelated within components, however correlation is induced between variables at the overall level by the mixing distribution. This can be seen as an extension of the local independence assumption for repeated measurements. This model has been called a “constrained” multivariate trajectory model (Nagin and Tremblay, 2001; Brame et al., 2001).

The third method, which has been called the “unconstrained” model (Nagin and Tremblay, 2001), decomposes the multivariate sample into a set of univariate mixtures, with separate mixing distributions. Using this method, each variable has a distinct set of components, so that each member of the sample has a set of posterior probabilities of class membership for each variable $\boldsymbol{\tau}_{jg} \quad (g = 1, \ldots, p)$. Variables are assumed to be independent of each other given their mixing distributions, but mixing distributions are not assumed to be independent. Association between the mixing distributions of the variables can be assessed by cross-classifying the modal class assignments, $\hat{\boldsymbol{Z}}_{jg}$, and using tests of association or log-linear models to analyse association. However, Nagin (2005) points out that tests of association carried out on contingency tables of $\hat{\boldsymbol{Z}}_{jg}$ can be biased, and will also underestimate standard errors, because they do not take account of classification error. He advocates simultaneously estimating all g mixtures using a likelihood based upon the joint mixing proportions $\pi_{g_1,g_2,\ldots}$.

The choice between the latter two models depends on practical considerations as well as the researcher’s hypotheses for the underlying model of the sample. Generally, the unconstrained model will provide a fit that is optimal for each variable, because the fitting algorithm does not have to compromise between the various variables. However, if the group structure on all variables is reasonably well represented by a joint set of groups, then the constrained model will tend to be better at distinguishing these groups, and will tend to have



(a) Constrained model (b) Unconstrained model

Figure 7: *Constrained and unconstrained group based trajectory models for multivariate counts.*

lower classification errors because there is more information available for the separation of the components, and more information available for the posterior classification of individuals. However, if the underlying group structure of the variables is disjoint, then the constrained model will tend to fit a large number of components. This is because the true joint group structure can only be represented by cross-classification of the components for each variable, with non-trivial proportions in the off-diagonal cells, and the constrained model will try to fit a component for each of the cells of this cross-classification.

Fitting the unconstrained model requires special software, or approximation using $\hat{\boldsymbol{z}}_{jg}$. In this study we will take the latter approach, although we are cognizant of the risk of bias this introduces due to classification errors.

It is not clear whether a single measure can be constructed upon which the two types of model can be compared. The unconstrained models are effectively being fitted in parts to different data, and because of the different parameters used in each we doubt that the log-likelihoods can be summed to create one measure. Therefore the strategy that has been adopted in this study is to compare the models on classification errors, size of components, and a tendency to find a large number of clusters.

4.b.ii Modeling overdispersion in counts

The usual model for counts, aggregated into time periods, is the Poisson model. In its simplest form, the group-based trajectory model is based upon a Poisson model, where the rate parameter is allowed to vary with age. However, it is commonly observed (Hinde and Demetrio, 2010) that the assumption of equal variance and mean implicit in the Poisson distribution imposes too much structure on real count data. If the variance is greater than the mean, this is called overdispersion.

To relax this strong assumption, various models have been used. To deal with overdispersion, the most common strategy is to allow the rate parameter itself to be a random variable, and to model the counts using a mixture distribution such as the negative binomial (Poisson mixed with gamma distribution for the rate parameter). In the negative binomial model component densities are defined as:

$$p_i(y_{jt}|\texttt{Age}_{jt}, \mathbf{z}_j, \boldsymbol{\theta}_i) = \frac{\Gamma(y_{jt} + 1/\alpha)}{y_{jt}!\Gamma(1/\alpha)} \left(\frac{1/\alpha}{1/\alpha + \mu_{it}}\right)^{1/\alpha} \left(\frac{\mu_{it}}{1/\alpha + \mu_{it}}\right)^{y_{jt}} \tag{1}$$

where α is the class-independent overdispersion parameter, and $\mu_{it} = \lambda_{it} = E[y_{it}]$ is the class-specific mean conditional upon age. This is the Latent Gold parametrisation. In R α is replaced with $\theta = 1/\alpha$. To avoid notational confusion since θ has already been defined, this quantity will be denoted by $\nu = 1/\alpha$.

Alternatives include the log-normal distribution, or a discrete (finite) mixing distribution, which essentially is the basis for a group-based trajectory model. These mixing distributions are sometimes combined, so for example a group-based trajectory model will model the counts, conditional upon group membership, as negative binomially distributed rather than Poisson distributed, to account for overdispersion after the groups have been taken into account.

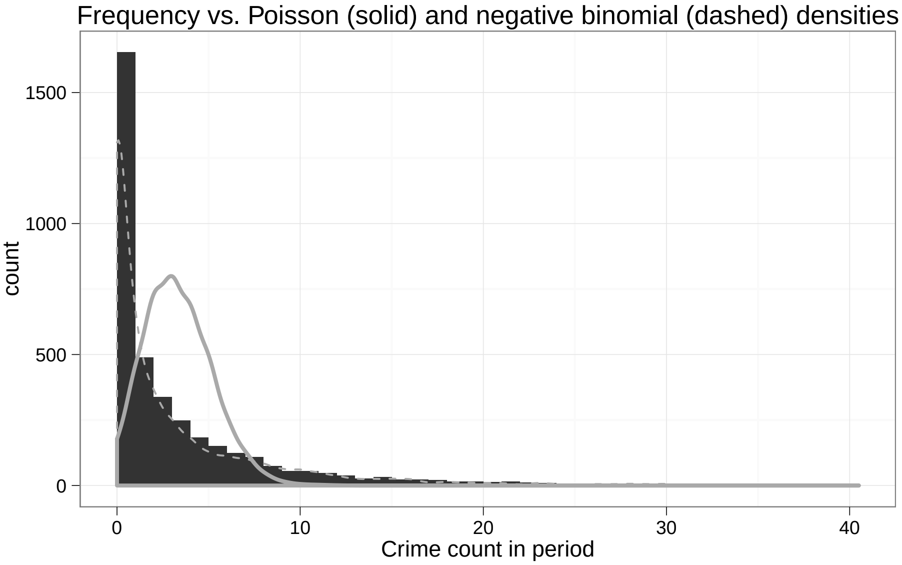


Figure 8: *The frequency of counts in pre-referral periods is compared to Poisson and Negative Binomial ($\nu = 0.37$) densities.*

Figure 8 compares the empirical distribution of counts to both a simulated Poisson and a simulated negative binomial distribution fitted to the data. It can be seen that the Poisson distribution seriously underestimates both the right tail of the distribution, and the number of zeroes. The Poisson mode is at the mean, whereas in the observed data the mode is at zero. In contrast, the negative binomial distribution fits the observed frequencies very well, but still underestimates the concentration of zeroes by about a quarter.

A negative binomial distribution has been adopted for the models in this Section, to account for overdispersion in the sample. However, it is clear that the adoption of the negative binomial distribution does not account for all of the excess zeroes. A class of models designed specifically to account for an excess of zeroes in count processes is the class of zero-inflated and zero-altered models.

4.b.iii Accounting for right-censoring and intermittency in the sample

In Section 2.b we outlined two types of missingness in the sample, one of which related to unrecorded periods of incarceration, and the other to non-ignorable right censoring at the point of referral.

The negative binomial model can be adjusted to account for the first of these by employing a zero-inflated negative binomial model. This is essentially a mixture model in which one of the components has a negative binomial distribution, and the other component is a point mass at zero.

$$p_i(y_{jt}) = \phi I_{(y=0)} + (1-\phi)p_{\text{negbin},i}(y_{jt})$$

where $p_{\text{negbin},i}(y_{jt})$ is the negative binomial distribution as defined in (1), ϕ is a class-independent zero-inflation weight, and:

$$I_{(y=0)} = \begin{cases} 1 & \text{if } y = 0, \\ 0 & \text{everywhere else.} \end{cases}$$

The zero-inflation weight, ϕ, in common with π in the general mixture specification, can take covariates that predict membership of the zero-inflated component, by the addition of a logistic model for the zero-inflation probability. We implemented the zero-inflated model with age, and three indicators of having been handed a custodial sentence in each of the last three periods, as predictors of zero-inflation.

The use of the zero-inflated model to account for unrecorded incarceration is motivated on the grounds that, if it is assumed that the counts of offenders who are free to offend follow a negative binomial distribution, then zero counts in excess of those predicted by the negative binomial must be caused by some other process. We concede that it is a leap of faith to assume that this process is incarceration, but the otherwise good fit of the negative binomial to the data (c.f. Figure 8) lends some empirical weight to the assumption.

In practice, unfortunately, Latent Gold was not able to accommodate the model described above, since the zero-inflated model implemented in Latent Gold treats zero-inflation as occurring at the level of sample member j, rather than observation jt. To circumvent this problem the zero-inflated model was fitted separately in R. The posterior probabilities of zero-inflation, τ_j^z were then calculated by:

$$\tau_j^z = \frac{P(z^{zi}=1)P(y_{jt}=0|z^{zi}=1)}{P(y_{jt}=0)} = \frac{\phi I_{(y=0)}}{\phi I_{(y=0)} + (1-\phi)p_{\text{negbin},i}(y_{jt})}$$

The posterior probability of non-zero-inflation, $\tau_j^{\neg z} = 1 - \tau_j^z$, was then added to the dataset. This probability was used as a weight on the observations in the group-based trajectory models. Reducing the weight of an observation y_{jt} to zero is equivalent to setting that observation to missing, so reducing the weight of an observation to $0 \leq \tau_j^{\neg z} \leq 1$ is equivalent to treating the observation as missing in proportion to the probability that the observation is zero-inflated.

Missingness due to right censoring at the point of referral is not unrecorded, but the fact that all sample members are right-censored before desistance introduces upwards bias in the tails of trajectories. Many of the offenders who were referred to the MTC would have been released before the end of the observation period, and although some of those would have committed more crimes, many would not. There was no easy way to alleviate this bias, but it could not be ignored because it had a sometimes-drastic effect on the fitted trajectories.

The approach taken in this study has been to extend the zero-inflated model to the post-referral period for each sample member. The effect of this is that observations are no longer deterministically missing after referral, but missing in proportion to τ_j^z. The justification for this method is that referral to the MTC is a type of incarceration for which the duration was not recorded, and does not take account of offenders who were subsequently released and who desisted from offending. It should be noted that in practice this method produces weightings that are very small for observations that would have been missing after referral.

4.b.iv Top-coding large counts

Weakliem and Wright (2009) analysed simulated data generated from an asymmetrical continuous mixing distribution, and found that finite mixture models tended to "find" multiple classes when there were none. They attributed this phenomenon to an excess of very large counts. In the MTC dataset, large counts tend to be associated with multiple counts of the same crime, either occurring on the same occasion, or more usually over multiple occasions, having been brought to justice on one occasion. There is arguably a point where one offender committing the same crime does not add more information to the sample, and risks exerting disproportionate influence on the shape of trajectories or probabilities of class membership.

For this reason, the counts of total offending have been re-coded, with any larger than 40 coded as 40. Similarly, counts for general offending and sexual offending have been top-coded at 40 and 30, respectively. Figure 9 shows the distribution of the counts with and without top-coding; only a handful of counts are affected.

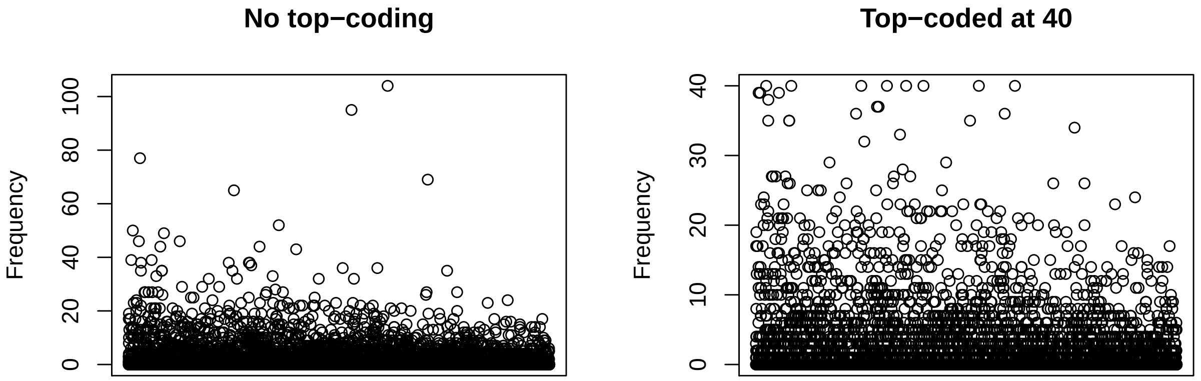


Figure 9: *Index plot of counts with and without top-coding at 40.*

4.b.v Cubic splines and the calculation of confidence bands

Trajectories in group based trajectory models are usually specified by using either quadratic or cubic polynomials to allow for curved trajectories. However, standard polynomials have two limitations that are important in the context of this study. Firstly, the fitting of a polynomial curve is "non-local" meaning that a data point in one part of the response space (observation period) can influence the shape of the curve in a distant part of the response space. This is a concern in this study because of the sparsity of observations at the end of the observation period, and the risk that the fitted trajectory in this portion of the observation period will be influenced by data points in the denser region of the period.

A second and, given the aims of the study, much more restrictive limitation is that neither a quadratic nor a cubic polynomial is capable of producing curves that are arbitrarily multimodal. A cubic polynomial can fit a curve with an "up-tick" at the end, but it cannot produce a curve that rises and then falls more than once. Since neither of these limitations are acceptable, in this study cubic splines have been employed as a flexible alternative to polynomials.

In its simplest form, a cubic spline with one knot h_1 can be fitted by adding a basis to the design matrix of a cubic polynomial (e.g. $X = (\mathbf{1}, \mathtt{age}, \mathtt{age}^2, \mathtt{age}^3)$). The additional basis has the form:

$$(\mathtt{age} - h_1)_+^3 = \begin{cases} (\mathtt{age} - h_1)^3 & \text{if } \mathtt{age} > h_1, \\ 0 & \text{if } \mathtt{age} \leq h_1. \end{cases}$$

Additional bases can be added to accommodate more than one knot point in the data.

Although simple to define, spline bases calculated in this way suffer from the fact that the basis functions are usually highly correlated with each other, which can lead to numerical instability and imprecision in the fitted estimates (Keele, 2008). To reduce this problem, a B-spline basis is an orthogonal transformation of a spline basis, that preserves all of the fitting properties of a simple spline whilst avoiding collinearity.

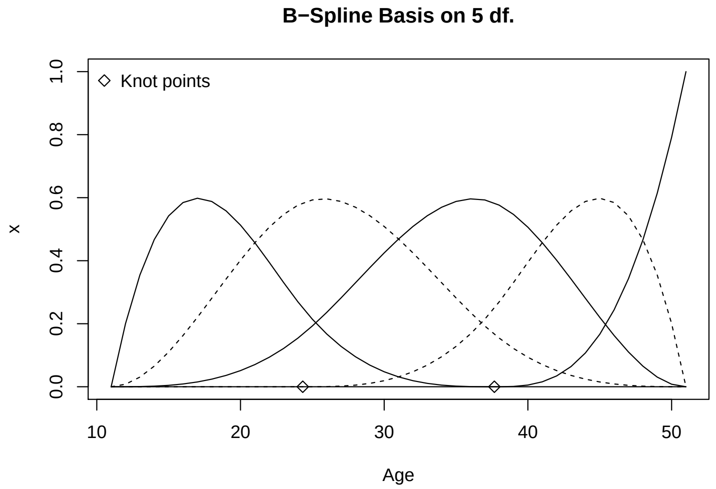


Figure 10: *Basis of a third degree (cubic) B-spline with two knot points. The degrees of freedom of the basis = degree + knots.*

Figure 10 shows the values of the basis functions used in the study, augmented to yearly (rather than five-yearly) periods. This augmented design matrix was used for plotting to provide a smooth interpolation of trajectories between five year mid-points.

95% pointwise confidence bands were calculated to provide a visual indication of the imprecision in the fitted splines. The 95% pointwise confidence interval for the vector of fitted values $\hat{\mathbf{y}}$ is calculated by back-transforming the confidence interval limits of the vector of linear predictors, $\hat{\boldsymbol{\eta}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, where $\mathbf{X}$ is the design matrix and $\hat{\boldsymbol{\beta}}$ is the vector of parameter estimates. The upper and lower limits of the confidence interval of $\hat{\boldsymbol{\eta}}$ are given by:

$$[L_{95\%}(\hat{\boldsymbol{\eta}}), U_{95\%}(\hat{\boldsymbol{\eta}})] = \hat{\boldsymbol{\eta}} \pm Z_{0.025}\,\mathrm{s.e.}(\hat{\boldsymbol{\eta}})$$

Where $Z_{0.025} \approx -1.96$ is the 2.5^{th} centile of the standard normal distribution, and $\mathrm{s.e.}(\hat{\boldsymbol{\eta}})$ is the vector of standard errors of $\hat{\boldsymbol{\eta}}$. This is given by:

$$\mathrm{s.e.}(\hat{\boldsymbol{\eta}}) = \mathbf{X}\,\mathrm{cov}(\hat{\boldsymbol{\beta}})\mathbf{X}^T$$

where $\mathrm{cov}(\hat{\boldsymbol{\beta}})$ is the covariance matrix of $\hat{\boldsymbol{\beta}}$.

The confidence interval limits thus calculated are then backtransformed using the inverse link function $g^{-1}(\cdot) = \exp(\cdot)$ to give the confidence interval limits of the fitted trajectory $\hat{\mathbf{y}}$.

4.b.vi Choosing the number of components

The choice of the optimal number of components in a GBTM, as in any finite mixture model, is an important consideration, and does not have a single well-defined solution. The number of components used obviously has a large bearing on the fit and on the interpretation of the model, but it is not determined by the model, and must be provided *a priori* by the researcher. A common approach is to fit a sequence of models with increasing numbers of components, and compare them using some summary statistic. However, this too is complicated because a set of finite mixture models are not nested, and so likelihood ratio tests are not available[5].

To deal with this problem, researchers often use one of the many available information criteria - measures that approximate the Kullback-Leibler information divergence, which is the information available to distinguish the modeled distribution from the empirical distribution of the data.

However, most information criteria are based upon the likelihood of the observed data, $\mathbf{Y}$, given the model. The observed data likelihood does not take account of the likelihood of the unobserved data, which in the case of finite mixture models is the unobserved set of class-membership indicators, $\mathbf{Z}$. The fact that this data is unobserved should mean that all possible values are equally likely. However, in the missing data conceptualisation of finite mixture models, it is assumed that the true $\mathbf{Z}$ is a collection of zeroes and ones, and cannot take any value in between. This means that any model that moves the estimates $\hat{\boldsymbol{\tau}_{ij}}$ away from zero or one, makes *any* possible values of $\mathbf{Z}$ *less* likely. Note that this distinction is only important if we are treating the "missing data" conceptualisation as true; if we do not believe in the existence of some unobserved vector of class memberships, then the observed-data likelihood is the same as the complete-data likelihood.

[5] Unless some form of bootstrapping is used, but such methods are not considered in this study.

This distinction between the observed-data likelihood and the complete-data likelihood (also called the classification likelihood) has given rise to the formulation of information criteria based upon the latter. These measures take advantage of the fact that, assuming independence, the joint likelihood of two partitions of the set of random variables can be found by the product of the likelihoods for each partition, or equivalently by the sum of their log-likelihoods. Therefore, the log-likelihood, $\log L(\mathbf{Y}, \mathbf{Z}) = \log L(\mathbf{Y}) + \log L(\mathbf{Z})$. The mean of the log-likelihood of $\mathbf{Z}$, conditional upon the observed data, can be approximated using the negative entropy of $\hat{\boldsymbol{\tau}}$, the vector of estimated posterior probabilities of class membership, where the entropy is given by:

$$\mathrm{EN}(\hat{\boldsymbol{\tau}}) = -\sum_{i=1}^{k}\sum_{j=1}^{n} \hat{\tau}_{ij} \log \hat{\tau}_{ij}$$

Biernacki et al. (2000) use this quantity to create a measure called the Integrated Classification Likelihood (ICL). Although the full form of the ICL is difficult to calculate, there is an approximation to the ICL that has been shown to have good properties if the size of clusters is large, called the ICL-BIC. The form of the ICL-BIC is as follows:

$$\text{ICL-BIC} = -2 \log L_{OBS} + 2\,\mathrm{EN}(\hat{\boldsymbol{\tau}}) + d \log N$$

where $\log L_{OBS}$ is the observed-data log-likelihood, d is the number of parameters fitted by the model, and N is the sample size. It can be seen that the ICL-BIC is equal to the BIC with the addition of $2\,\mathrm{EN}(\hat{\boldsymbol{\tau}})$. In practice, this extra quantity is easily calculated using the posterior probabilities output by Latent Gold.

The ICL-BIC has been shown to perform well in simulations at finding the true number of classes (McLachlan and Peel, 2000) and is less likely to overfit than the BIC or many other information criteria. One situation where discretion might be needed is when the ICL-BIC indicates that the optimal number of groups is one. Since a single component model is bound to have an entropy of zero this is a risk, and it is not entirely clear whether the use of a criterion based upon the classification likelihood should be taken to its logical conclusion in such a situation. If this situation arises the one-class model will be reported as the most parsimonious model, and if there are other local minima for models with more than one component they will also be investigated.

Since this study is not being conducted using Bayesian methods, there is no way to incorporate a prior belief on the number of components. However, this does not mean that we do not have prior beliefs. We believe that there is not necessarily any objective truth to the idea of classes of trajectory, but that classifying criminal careers into a small number of trajectories is a useful way to begin to understand and interpret them. With these purposes in mind, our "prior" for the number of classes is that it is greater than one, and not so great that, when trajectories are plotted, they look like spaghetti.

It is fair to ask whether a criterion that maximises the complete-data likelihood is desirable, given that we have expressed some skepticism about the existence of unobserved classes in the data. However, it is still true that the main methods of analysis of, and interpretation of GBTMs are through the window of either plotted trajectories, or modal class assignments. The ICL-BIC, tending to prefer well-separated classes and smaller numbers of classes, is well-suited to the purpose of producing easily interpretable, stable groups. If the purpose of the model were to estimate the trajectories of individuals in the study, such as was done by Blokland et al. (2005), then it would be preferable to employ an information criterion that maximised the observed-data likelihood and fit as many components as necessary, as they point out.

4.c Method of fitting the total frequency and dual frequency models

All of the group based trajectory models were fitted in Latent Gold (Vermunt and Magidson, 2005) using the Expectation-Maximisation algorithm, with a Newton-Raphson step to aid rapid convergence. The counts were combined with a B-spline design matrix, the exposure indicators and the zero-inflation weights. The total frequency models were also fitted with weights equivalent to censoring all of the observations after referral, and with no weighting, equivalent to no censoring, for comparison.

Both constrained and unconstrained variants of the GBTM were fitted. The two types of model were compared for classification performance, cluster sizes, and tendency to fit a large number of groups, favouring those where both classification errors and the number of groups tended to be small. We preferred the smallest class size not to be much smaller than 10% due to the risk that the effective class size at the later ages would become very small. Each model was fit with one-five classes, and the optimal number of classes within each model/sub-model was determined using the ICL-BIC measure.

4.d Results of the total frequency models

4.d.i Results of the zero-inflation model

The effective sample sizes resulting from the use of zero-inflation weights, as well as the effective sample sizes from both full right censoring and no right censoring are presented in Table 11. It can be seen that the effective sample sizes for the zero inflation and fully censored methods are similar, but that the zero inflation method increases the effective sample size at the end of the observation period.

4.d.ii Results of the group based trajectory models

Table 6 shows fit statistics for the models on total frequency, fit with one-five classes, and the model fit to two classes using the alternative methods of accounting for missingness. Using BIC, the model with three classes would have been optimal. However, the ICL-BIC indicates that the model with only one class is the optimal model. Apart from the one-class model, the lowest ICL-BIC is for the model with two classes. All of the models with ZI coding, apart from the one-class model display large classification errors, whereas the classification errors from the censoring and no-censoring methods are less severe.

Figure 12 shows a comparison of the fitted trajectories from the three methods of censoring observations. It can be seen that the approach where observations are right censored after referral (fully censored) produces trajectories that do not decline with age. The fully-censored

Figure 11: Effective sample size by age for the three methods of dealing with missingness. Table also shows ratio of ZI to censoring method.

(a)

Age	Zero-inflation	Censored	Not censored	ZI:C
11	821	821	821	1.00
16	733	733	733	1.00
21	582	593	672	0.98
26	470	425	701	1.11
31	317	292	753	1.09
36	205	185	771	1.11
41	113	108	792	1.04
46	61	63	806	0.98
51	47	37	811	1.26

(b)

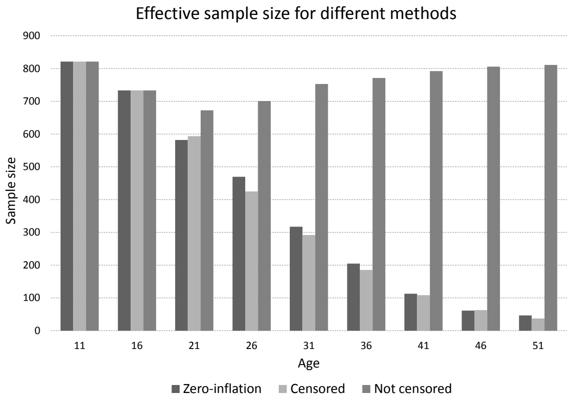


Table 6: Fit statistics for the models on total frequency

No. Cl.	LL	BIC	Npar	Entropy	ICL-BIC	Class. Err.	Max. π	Min. π
Models using zero-inflation weighting								
1	-8551	17150	7	0	17150	0.000		
2	-8402	16906	15	264	17434	0.151	0.52	0.48
3	-8370	16894	23	442	17778	0.249	0.49	0.23
4	-8349	16906	31	553	18012	0.311	0.42	0.12
5	-8332	16925	39	557	18039	0.315	0.41	0.05
Model using full censoring								
2	-8222	16544	15	186	16916	0.107	0.54	0.46
Model using no censoring								
2	-9750	19602	15	118	19838	0.053	0.66	0.34

data also produces trajectories with larger standard errors in the later ages than the other two methods. This is to be expected, since the size of the confidence bands is directly related to the effective sample size. The confidence bands of the trajectories for the zero-inflation method overlap more than the trajectories for the other two methods, which probably explains why the classification was less well-separated. This is not necessarily a fault of the method, since the fully censored method might have been expected to have produced even worse classification errors, had the trajectories been closer, because of the smaller effective sample size as evidenced by the wider confidence bands.

The fully censored method and the zero inflation method produce groups that are similar in size and in general shape. In contrast, the uncensored data produces trajectories that are completely different. In particular, the order of group size is reversed between the adolescent-onset and accelerator groups.

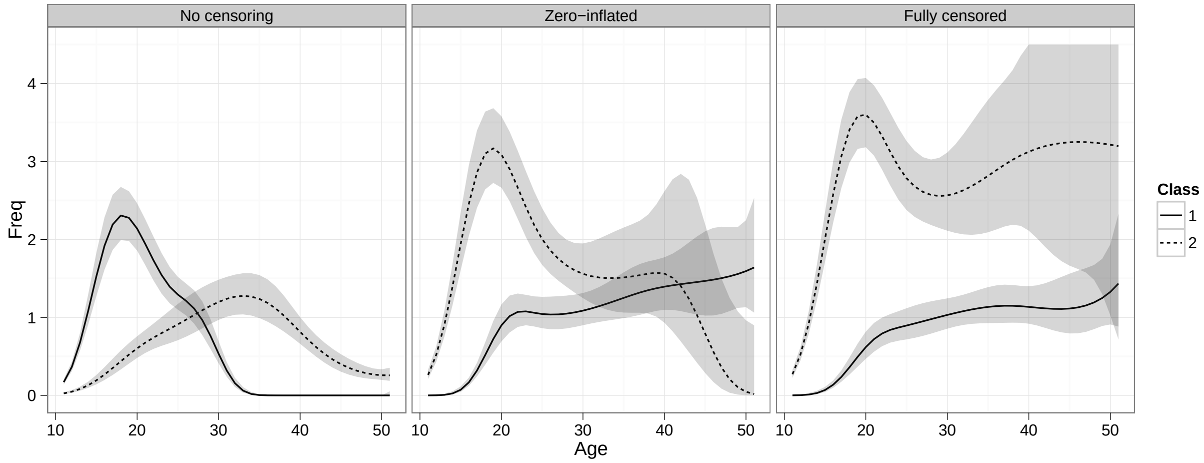


Figure 12: *Comparison of three different methods for treating missingness. Two-class models for total frequency.*

4.e Evaluation of results of the total frequency model

The zero inflation model and the fully censored model both produce reasonably similar results in terms of component size, and the fitted trajectories are similar, apart from the tendency of the fully censored data to produce trajectories that do not decrease, and sometimes increase without bounds in the later ages. Although only one comparison is shown, this tendency was observed to be consistent in a number of other models, the observation of which largely motivated the use of the zero-inflation weights to augment the effective sample size at the later ages. The unorthodox use of a zero-inflated model might be considered a trick to bring the trajectories down to a reasonable level, and is justified only as a crude way of counterbalancing the bias caused by non-ignorable missingness explained in section 2.b.

Either model shows some evidence of bimodality in at least one of the trajectories. However, the second peak, where present, seems to be much later than predicted by Hanson (2002) or Lussier (2010), who predict the second peak at around the age of 30. In fact, at around the age of 30 there appears to be a slight lull in offending, and the second peak is after the age of 40. Only the uncensored data produces a model that agrees with Hanson or Lussier.

In the next section, the frequency of criminal activity will be disaggregated into sexual and non-sexual crimes, to investigate whether this bimodality can be explained by crime mix.

4.f Results of the dual frequency (sexual/general) models

Table 7 presents the fit statistics resulting from fitting the constrained models with one-five classes. Table 8 presents the fit statistics for the unconstrained models The classification errors are lower in the constrained models than in either of the sets of unconstrained models. Again, the minimum of the ICL-BIC is at the one-class constrained model. ICL-BIC was not

Table 7: Fit statistics for the constrained models for sexual and general offending.

No. Cl.	LL	BIC	Npar	Entropy	ICL-BIC	Class.Err.	Max. π	Min. π
1	-11920	23927	13	0	23927	0.000		
2	-11692	23565	27	208	23981	0.109	0.65	0.35
3	-11569	23413	41	274	*23961*	0.135	0.50	0.24
4	-11521	23412	55	343	24098	0.165	0.40	0.10
5	-11442	23348	69	354	24056	0.171	0.38	0.09

Table 8: Fit statistics for the unconstrained models for sexual and general offending.

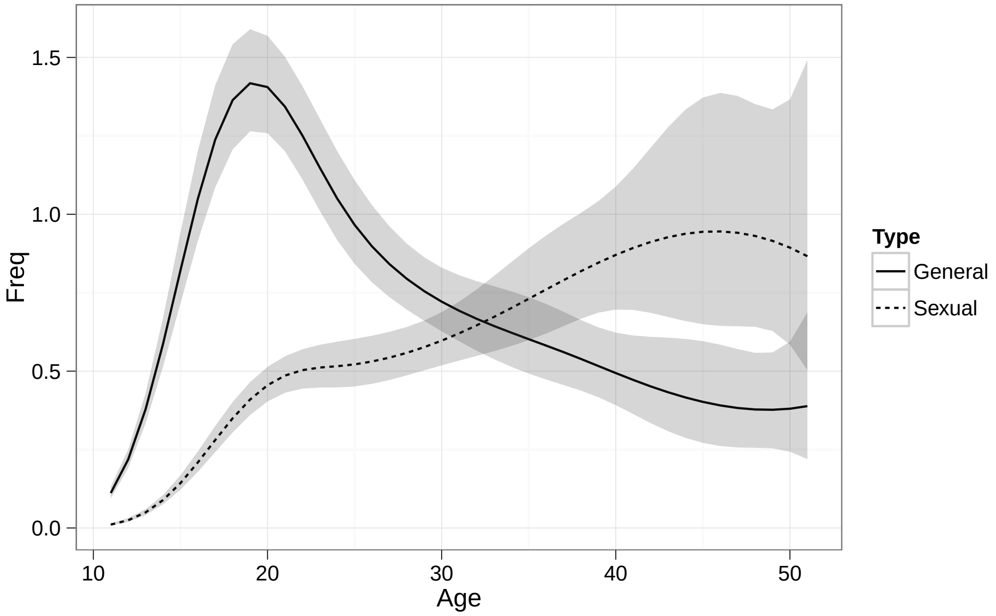
No. Cl.	LL	BIC	Npar	Class.Err.
Unconstrained models for general offending				
1	-6946	13938	7	0.000
2	-6816	13733	15	0.142
3	-6766	13686	23	0.226
4	-6747	13702	31	0.278
5	-6730	13722	39	0.361
Unconstrained models for sexual offending				
1	-4974	9995	7	0.000
2	-4925	9951	15	0.277
3	-4911	9976	23	0.344
4	-4910	10028	31	0.485

calculated for the unconstrained models, but due to the large classification errors might be expected to give the same minimum at one class. The trajectories for the one-class model are plotted in Figure 13. The trajectory for general offending is akin to a classic age-crime curve, whilst the trajectory for sexual offending has a peak in the mid-40s. The shapes of the two trajectories for general and sexual offending are very similar to the shapes of the first and second classes in the model for total frequency.

Figure 14 shows the three-class model for sexual and general offending. The first set of plots show the complete trajectories; the confidence bands, particularly for the second class (26%) become very large after the age of 40. This is presumably because the effective sample size for this group becomes very small (perhaps zero) after this age. The confidence bands for the other trajectories are similarly wide. This demonstrates that even clusters that are reasonably large at the beginning of the period might have very small effective sample sizes towards the end of the period. In the second set of plots the trajectories have been truncated at the age of 40 to more clearly show the unaffected parts of the trajectories.

The first class (50%) has a general offending trajectory that peaks at a low rate in the mid-20s and then declines, but not to zero. This class has a sexual offending trajectory that rises steadily from the age of 15 to middle age. The second class (26%) has a general offending trajectory that can best be described as “adolescent-limited” and a similarly limited sexual offending trajectory. The third class (24%) has a general offending trajectory that peaks slightly later, just after the age of 20, and declines slowly, but is still greater than one offence per year into middle age. The sexual offending trajectory for this class is similar to the sexual offending trajectory for the second class, and in fact is indistinguishable, based on the confidence bands. Interestingly, the BIC for the unconstrained models indicate minima at three and two class models for the general and sexual offending, respectively.

Figure 13: *Trajectories for general and sexual offending. One-class constrained model.*



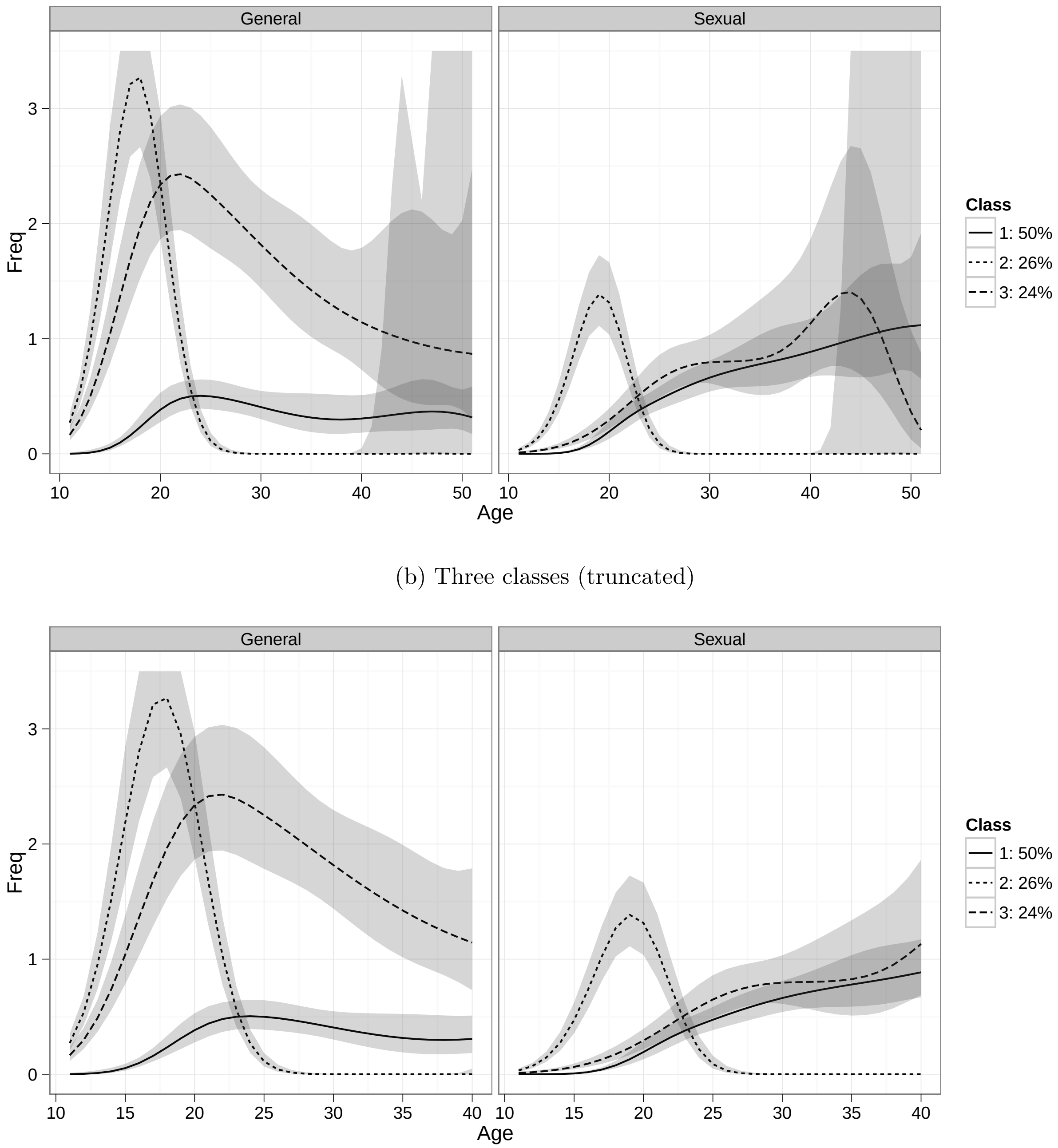
4.g Evaluation of results of the dual frequency model

There is strong evidence of bimodality in the one-class dual frequency model, with each offender expected to have a peak of general offending in late adolescence, and a peak of sexual offending after the age of 40 (c.f. Figure 13). This model is equivalent to a fixed-effects spline regression. The shape of the fitted trajectories echoes the shapes of the two-class total frequency trajectories. This suggests that the two-class model for total frequency in the last section can be partly explained by the relative weighting of different crime types in the criminal profiles of offenders.

The three-class mixture reinforces the evidence of bimodality and late onset of sexual offending: three quarters of offenders are classified on trajectories of sexual offending that increase steadily into middle age. The three quarters of offenders whose sexual offending increases with age can be further split into those who commit relatively few non-sexual offences over a long time period, and those who commit a large number of non-sexual offences over a long time period.

The other quarter of offenders are adolescent-limited in both general and sexual offending. This group is truly adolescent-limited, in that the frequency of both types of offending reduces

Figure 14: *Trajectories for general and sexual offending. Three-class constrained model, with and without truncation at 40.*



(a) Three classes

(b) Three classes (truncated)

to zero by the mid-20s. Such a group of strictly adolescent-limited *sexual* offenders is not predicted, as far as we know, by any previous studies or theories.

It can already be seen, with the one-class dual frequency model, that the disaggregation of crimes explains some of the apparent variation in trajectories at the aggregate level. The next stage of the analysis will investigate whether the classification of crimes into sexual and non-sexual itself obscures heterogeneity in trajectories at a lower level.

5 Analysis Three: Poisson-Log Normal Factor Trajectory Analysis

5.a Motivation

Although it is an improvement on the single frequency model, the dual frequency model nevertheless still relies on an *a priori* grouping and aggregation of the crimes. This aggregation might conceal differences in the trajectories of different crimes within the group. It is possible to continue to disaggregate into smaller and smaller groups; however with each subdivision the models become harder to fit and harder to interpret.

In order to model the trajectories in a way that is parsimonious, without hiding potentially different trajectories, it is desirable to find a way to partition the set of crime types into groups that are guaranteed to have trajectories that are as close to proportional as possible. For this purpose it is sufficient to find groups of crimes whose frequencies of occurrence are as positively correlated as possible within each time period, because if crimes are positively correlated within each time period, then their trajectories will be proportional, and the within-t correlation between the frequencies of any two types of crime is a measure of the degree to which their trajectories will be proportional.

To this end we introduce a method of dimension reduction that explains the maximum possible within-t covariance in the sample in a set of continuous factors. We will show that, due to this property, this factor analytic model is useful for the analysis of multivariate criminal careers, independently of the use of factors to hypothesise the existence of theoretical constructs.

5.b Factor analytic models for count data

A factor analytic model is a model for dimension reduction of multivariate data. It is based upon the principle that the dependence structure of a set of observed variables can be represented by a smaller number of latent continuous variables. To formulate the model, the observed or manifest variables, $\mathbf{Y}_g \quad (g = 1, \ldots, p)$ are assumed to be independent of each other, given a smaller set of unobserved random variables, $\mathbf{F}_l \quad (l = 1, \ldots, q; q < p)$. It is assumed that the distribution of these unobserved variables is multivariate normal, and that the $\mathbf{F}_l$ are related to the $\mathbf{Y}_g$ by a linear measurement model.

In the case that the manifest variables are all Poisson distributed random variables, the linear measurement model is linked to the rate parameter, λ_g, of the Poisson distribution of each variable, $\mathbf{Y}_g$, by a log link so that:

$$\log(\lambda_g) = \mu + \sum_l \Gamma_{kl} f_l$$

where Γ_{kl} are the loadings of factor f_l on each observed g. Since a linear combination of normals is also normal, this model produces a lognormal mixing distribution for the rate parameter, λ of each observed variable. The shape of the lognormal distribution is similar to the shape of the gamma distribution, so the lognormal factor model allows for overdispersion in the observed variables in a way that is reasonably interchangeable with using negative binomials. The loadings, analogous to regression co-efficients, encode the relative weight of each factor in creating the λ of each manifest variable.

The distribution of $\mathbf{Y}$ obtained by integrating over the joint distribution of $\mathbf{F}$:

$$p(y_{jgt}|\boldsymbol{\Omega}) = \int \prod_{g=1}^{p} p(y_{jgt}|f_{jlt}, \boldsymbol{\Gamma}) f(f_{jlt}|\boldsymbol{\Upsilon}) \, \mathrm{d}f_{jlt}$$

where $\boldsymbol{\Upsilon}$ are the means and covariance matrix for the prior of the multivariate normal factors, and $\boldsymbol{\Omega}$ contains all parameters of interest.

This model can be seen as a generalisation of the Poisson-log normal multivariate model of Aitchison and Ho (1989), except that in the latter each observed variable is accompanied by a unique log normal variable, and there are no cross-loadings. Apart from the benefits of dimension reduction and interpretative simplicity, a practical advantage of reducing the number of log normal variables in the mixing distribution is that fitting the model becomes computationally less difficult to estimate because there are less dimensions to integrate over.

In contrast with normal linear factor analysis, estimates of the "uniquenesses", residual variance unique to each manifest variable, are not calculated, because it is assumed that the variance of the Poisson variables is equal to the mean. It is possible to calculate these unique variances, but this would require calculating the full p dimensional covariance matrix, which would entail integrating over p, rather than q, dimensions. The lack of residual variance in the mixing distribution means that factors tend to load very highly onto at least one manifest variable, because they are maximising the explanation of total excess variance (due to overdispersion) rather than covariance, in a way very similar to probabilistic PCA (McLachlan and Peel, 2000, p.149) and interpretation of the results should take this into account.

Factor scores are calculated in MPlus (Muthen and Muthen, 2011) by maximising the log of the posterior distribution of $\mathbf{f}_{jt}$ given the observed $\mathbf{y}_{jt}$:

$$g(\mathbf{f}_{jt}|\mathbf{y}_{jt}) \propto p(y_{jt}|f_{jt}, \boldsymbol{\Gamma}) f(f_{jt}, \boldsymbol{\Upsilon})$$

Identifiability is an important concern in factor analysis. The log normal factor model is similar to linear factor models in being subject to three types of invariance: location invariance; scale invariance; and rotation invariance (Bartholomew and Knott, 1999). The first two of these imply that the distribution of factors is undefined, and must be fixed by setting a prior on the distribution of the factors. In order for the model to be identified, either the scale of the factors (the variance) or the location must be fixed. The scale can be fixed by specifying that the factors must be standard normal with unit variance. Alternatively the location can be fixed by constraining one of the loadings, $\boldsymbol{\Gamma}$ to one, thus tying the location of the factor to the location of the corresponding manifest variable. In this study, the prior is multivariate normal and the locations have been fixed by setting one loading on each of the factors to one. Rotation invariance means that it is not possible to distinguish empirically between a model with loadings $\boldsymbol{\Gamma}$, and a model with loadings $\boldsymbol{\Gamma}\mathbf{M}$, where $\mathbf{M}$ is a non-singular transformation matrix. Under such a transformation the factors would be transformed to $\mathbf{F}' = \mathbf{M}^T\mathbf{F}$.

The maximum number of degrees of freedom that can be used in a factor model is equal to $(p(p+1))/2$, which in the case of seven manifest variables is 28.

Exploratory and confirmatory factor analysis proceed in the same way as for linear factor analysis: in the exploratory phase no loadings are constrained to zero. Models with an increasing number of factors[6] are fitted by maximum likelihood. The model with the lowest value of BIC is selected. The factor loadings are rotated using oblique rotation. The rotated loadings that are close to zero (see Section 5.c.i) in the exploratory analysis are constrained to zero in the confirmatory analysis. These constraints usually improve identification of the model and remove rotational invariance.

5.b.i The interpretation of the Poisson log normal factor model

Aitchison and Ho (1989) justify the Poisson log normal model as a model for multivariate data by envisaging circumstances in which the observed count processes are dependent upon other, unobserved processes, and the count processes are independent of each other, conditional upon the unobserved processes, which are not. The example given by Aitchison and Ho (1989) was to do with butterfly and plant species, but it is easy to apply similar reasoning to criminal careers.

The occurrence of crime can be conceptualised as depending upon both the desire to commit a crime, and an opportunity to commit a crime. Crime opportunities (victims in the case of person crimes, unguarded property in the case of property crimes) are random events that can be assumed to be independent of each other, given the offender, whereas the offender's desire or propensity to commit a particular type of crime is not independent of propensity to commit other crimes, or of propensity to commit the same crime in adjacent periods. These propensities can be conceptualised as a set of random variables of dimension p, or they can be thought of as being fewer in number than the number of different crime types, or there might be only one (c.f. Gottfredson and Hirschi (1990)). This is not to say that, given the existence of such a set of propensities, the solution of the factor model is necessarily a faithful representation of it. However, it is useful to have such a conceptual model in mind because it clarifies the roles of the different parts of the model.

5.c Method

5.c.i Estimation of the factor measurement model

The estimation of the exploratory and confirmatory factor analysis models and the calculation of factor scores were carried out in MPlus (Muthen and Muthen, 2011). In MPlus both types of model were estimated by maximum likelihood with robust standard errors, using an E-M algorithm with numerical integration on seven points of support.

The exploratory factor analysis was carried out using one to three factors. Three factors was the maximum that could be fitted using degrees of freedom ≤ 28. Oblique rotation was carried out using the Geomin method.

[6] Limited by the total degrees of freedom available.

Based on the results of the exploratory factor analysis, a confirmatory factor model was constructed. All observed variables with scaled loadings greater than 0.5 were included in the measurement model for the corresponding factor. Cross loadings were allowed in a few cases, with a more stringent criterion of 0.6 for inclusion in the second measurement model.

5.c.ii Fitting the trajectory model to the predicted factor scores

Factor scores were calculated in MPlus and were exported into a text file for subsequent analysis in Latent Gold (Vermunt and Magidson, 2005). They were combined with a B-spline design matrix and the zero-inflation weights. Although the continuous factors are not zero-inflated, they do contain a large number of very low values that represent zeroes in the count variables. The effect of the zero-inflated model was to reduce the weight of these low values, and was equivalent to applying the zero-inflation weights to the observed variables themselves, had this been possible in MPlus.

The model fit in Latent Gold was a mixture of gaussian linear regressions with repeated values. Both constrained and unconstrained variants of the GBTM were fitted. The two types of model were compared for classification performance, cluster sizes, and tendency to fit a large number of groups, favouring those where both classification errors and the number of groups tended to be small. We preferred the smallest cluster size not to be much smaller than 10%. Each model was fit with one to seven classes, and the optimal number of classes within each model/sub-model was determined using the ICL-BIC measure.

From the unconstrained model, the modal class assignments $\hat{\mathbf{Z}}$ were cross-classified and used to construct joint probability contingency tables, which were then subjected to a Poisson log-linear analysis to determine associations between class assignments on the different factors.

5.d Results

5.d.i Estimation of the factor measurement model

Exploratory factor analysis was carried out on one-three factors. The solution with the lowest BIC was the three-factor solution. The loadings for the two- and three-factor rotated solutions are shown in Table 9, where a comparison of the two solutions reveals that there is a factor associated with rape, assault and property crimes, which in the three-factor solution is sub-divided into a factor weakly loading on to rape and assault, and including property crimes, and a factor loading onto rape, assault and non-contact sexual offences. In both solutions, there is a factor associated only with child molestation and non-contact sexual crimes. The underlined loadings are those where the two decision rules (loading greater than 0.5, or greater than 0.6 for cross-loading) were met.

Based on the results of the three-factor solution, a confirmatory factor model was constructed with a reduced set of loadings, and fitted. The loadings were included based on the criteria outlined in the method. The standardised results of the confirmatory factor model are presented in Figure 15. The standardised loadings of Factor one onto Property, Breaking and Entering and Other are very close to 1, which illustrates the tendency of the count factor analysis model towards high loadings. The loading of factor one onto rape is the lowest of the loadings of factor one. Factor two loads most strongly onto rape (0.83) and almost as

Table 9: Factor loadings from the two- and three-factor exploratory factor models

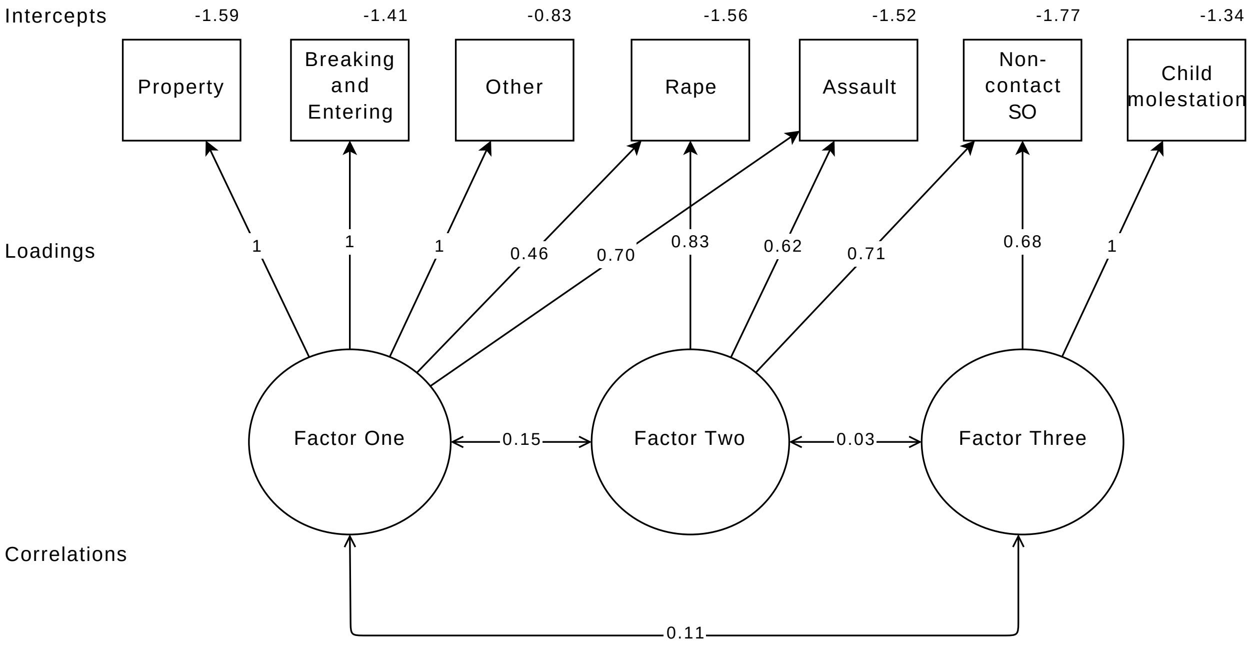
(a) Two Factor (BIC: 42,732)

Factor	1	2
Assault	1.000	0.003
Rape	0.989	-0.150
Child	0.138	0.990
Non-contact	0.410	0.912
B & E	0.964	0.265
Property	0.969	0.247
Other	0.893	0.449

(b) Three Factor (BIC: 39,694)

Factor	1	2	3
Assault	0.602	0.780	0.172
Rape	0.764	0.646	0.000
Child	-0.223	0.214	0.951
Non-contact	0.035	0.697	0.716
B & E	0.995	0.007	0.099
Property	0.988	0.101	-0.116
Other	0.811	0.007	0.585

Figure 15: *Fitted confirmatory factor model, with standardised intercepts, loadings and correlations.*



strongly onto non-contact sexual offences (0.71). Factor three loads only onto child molestation and non-contact sexual offences. The correlations between the factors are weak, and the correlation between factors two and three is especially small, which is perhaps surprising because the two factors both load onto non-contact sexual offences. All free parameters were significant based on Z scores.

5.d.ii Fitting the trajectory model to the predicted factor scores

The factor scores for each observation on the three factors were exported and used in constrained and unconstrained GBTMs. Tables 10 and 11 show fit statistics for the constrained and unconstrained models, respectively. It can be seen that the constrained model tends towards a large number of clusters, and the classification errors are in general slightly larger than those for the unconstrained model. The unconstrained model was favoured for these reasons.

In both types of model it can be seen that BIC consistently favours models with poor classification properties and a large number of classes. Indeed, in most cases BIC did not seem to reach a minimum by seven clusters, which was the largest number of groups considered. Especially in the unconstrained model, the ICL-BIC favours much smaller number of clusters than the BIC. Using the ICL-BIC, there are clear minima at three and two clusters for the first two factors. The ICL-BIC for the third factor has a minimum at two clusters, but the three cluster model has an ICL-BIC that is only fractionally (0.2) larger. As with any information criterion, which are asymptotically valid approximations of the Kullback-Leibler distance, such a small difference should not be considered decisive.

Four sets of trajectories are plotted in Figures 16, 17, 18 and 19; one set for each of the first two factors, and both the two-class and three-class trajectories for Factor 3. It can be seen that the main difference between the two- and three-class solutions for factor three is the addition of a third class, whilst the other two classes remain almost unchanged in shape. The largest class in both models (73%; 65%) has an almost linear, constantly rising trend into middle age. The second class (27%), which is completely unchanged in both models, is a class of participants who were not active in this factor. The size, π of the third class in the three-class model is smaller than 10%, and the average τ for sample members modally allocated to the third class is only 0.77 (compared to greater than 0.99 for the other two classes). However, the posterior probability of class membership is a function of π as well as the observed data, so smaller clusters can be expected to have smaller posterior probabilities. Whether to favour the two- or three-class model seems to be a matter for discretion. The extra trajectory in the three-class model is substantively interesting because it represents a group who were actively involved in child molestation and non-contact sexual offences during adolescence, and who then had a second, smaller peak of the same kind of offending in adulthood. The class represents around 8% of the sample, or around 60 participants.

Figure 16 shows the fitted trajectories for Factor One, with 95% confidence bands. Interestingly, since this is the factor with the strongest loadings for non-sexual offending, the largest class (60%) is represented by a trajectory that most resembles the classic age-crime curve, with a peak before the age of 20 and a slow decline. The second class (29%) is represented by a trajectory that peaks in the mid 20s and in the late 40s. In both of these classes the confidence interval around the fitted trajectory flares in the later ages, which indicates that the sample size available to fit these trajectories was much smaller towards the end of the time period. The third class (11%) is a class of participants who were not active in this factor.

Figure 17 presents the fitted trajectories for Factor Two. The confidence intervals around both of these trajectories are quite wide, indicating uncertainty in the fitting of the model. The classification errors for this model were also the largest of all of the unconstrained factor models. The largest class (67%) has a fitted trajectory with a peak in the mid-20s. There appears to be a second peak or rising trend beginning at around the age of 40, although the confidence bands after this point would also admit a flat trajectory from this point on. The second class (33%) is associated with a low trajectory that has no peak in the mid 20's, and rises steadily to around 40.

Table 10: Results for Constrained three-factor models

No. Cl.	LL	Npar	Entropy	BIC	ICL-BIC	Class.Err.	Min. π
1	-19740	19	0	39607	39607	0.000	1
2	-19320	39	157	38902	39216	0.085	0.42
3	-19031	59	193	38458	38844	0.101	0.20
4	-18771	79	191	38071	38453	0.095	0.12
5	-18550	99	174	37764	38112	0.082	0.11
6	-18351	119	196	37502	37894	0.090	0.12
7	-18218	139	231	<u>37369</u>	<u>37831</u>	0.102	0.10

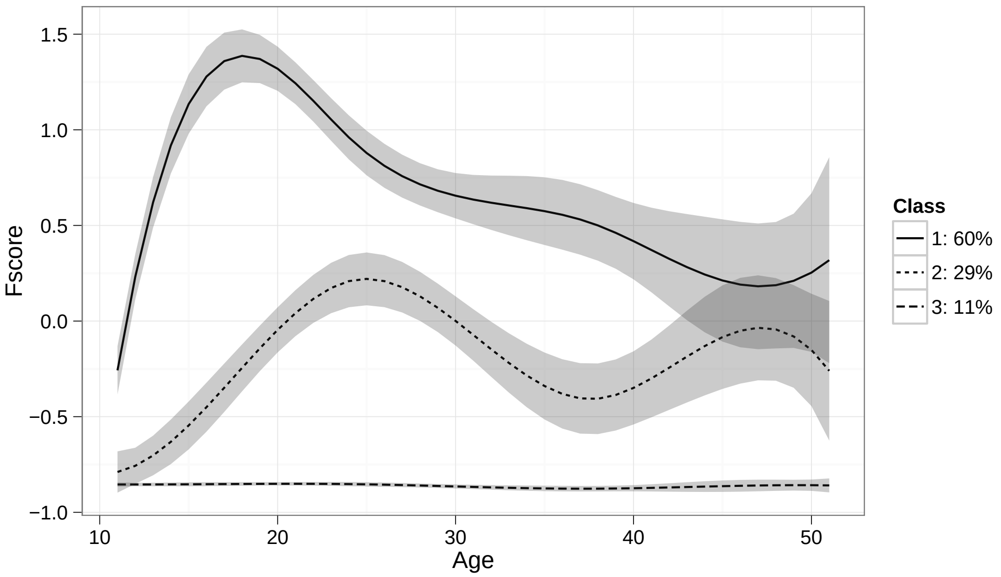


Figure 16: *Trajectories of factor one - two-class unconstrained model*

5.d.iii Fitting the log-linear model to cross-classified classes

Table 12 shows the three-way cross-tabulation of participants by modal class assignment. It can be seen that there are far more observations in the active (i.e. higher frequency) first classes for each of the factors. The appearance of a strong diagonal in this cross-tabulation would provide evidence to support a constrained model, in which classes are shared by all factors; this is clearly not the case, since the counts in the diagonals for all two-way

Table 11: Results for Unconstrained three-factor models

No. Cl.	LL	Npar	Entropy	BIC	ICL-BIC	Class. Err.	Min. π
Factor One							
1	-6567	7	0	13181	13181	0.000	1.00
2	-5508	15	0	11117	11117	0.000	0.11
3	-5308	23	169	10769	11107	0.091	0.11
4	-5245	31	346	10699	11391	0.172	0.11
5	-5198	39	354	10659	11367	0.173	0.07
6	-5166	47	388	10648	11424	0.180	0.06
7	-5134	55	471	10637	11579	0.221	0.07
Factor Two							
1	-4212	7	0	8471	8471	0.000	1.00
2	-3814	15	120	7729	7969	0.060	0.32
3	-3745	23	252	7645	8149	0.135	0.14
4	-3700	31	317	7608	8242	0.157	0.06
5	-3667	39	402	7596	8400	0.196	0.07
6	-3639	47	493	7594	8580	0.237	0.06
7	-3611	55	537	7591	8665	0.251	0.06
Factor Three							
1	-7626	7	0	15299	15299	0.000	1.00
2	-4957	15	0	10015	10015	0.000	0.27
3	-4881	23	50	9916	10016	0.028	0.08
4	-4828	31	93	9863	10049	0.046	0.05
5	-4783	39	224	9829	10277	0.111	0.04
6	-4737	47	225	9791	10241	0.110	0.01
7	-4726	55	263	9820	10346	0.134	0.01

interactions (F1:F2, F2:F3, F1:F3) are small in comparison to the counts in the off-diagonals. The largest cell overall corresponds to the highest-frequency class in all Factors.

The interaction terms from the Poisson log-linear model are shown in Table 13. The interaction terms in a log-linear model are good approximations to the log of the cross-product ratio between two levels of two cross-classified categorical variables, as long as the table is not sparse, and the table is structurally complete (i.e. it is possible for all cells to have non-zero values). There is only one empty cell in the contingency table, and there is no reason to believe it is structurally zero, so the Poisson log-linear model is appropriate. The log cross-product ratio takes positive values if the two levels are positively associated (i.e. membership of one class makes membership of the other more likely) and negative values if the two levels are negatively associated.

Before fitting the log-linear model each of the variables (modal class assignments) was re-levelled so that the reference category was the inactive or least active class. Because of this, a positive interaction effect is associated with positive correlation between the two factors involved in the interaction, and a negative interaction effect is associated with negative

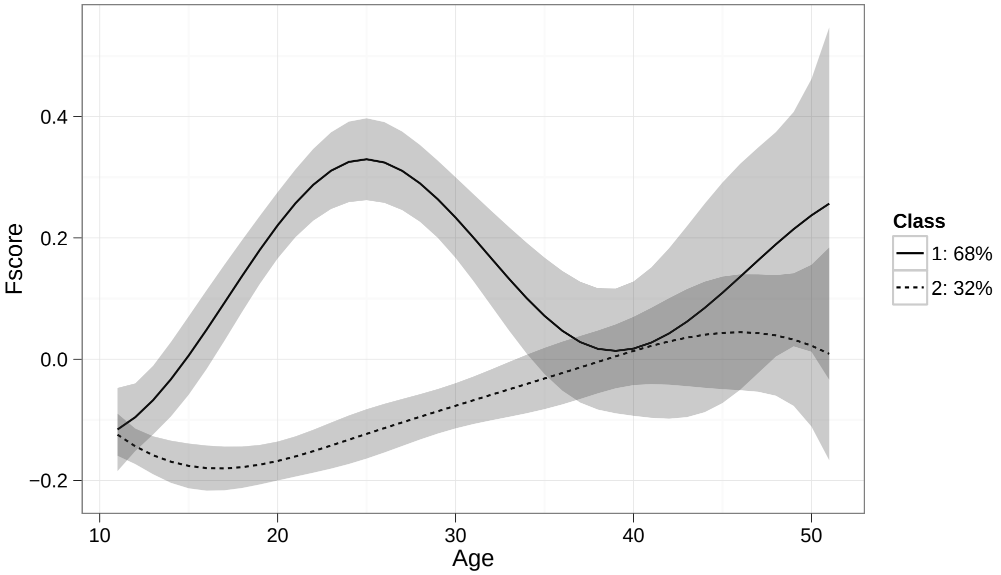


Figure 17: *Trajectories of factor two - two-class unconstrained model*

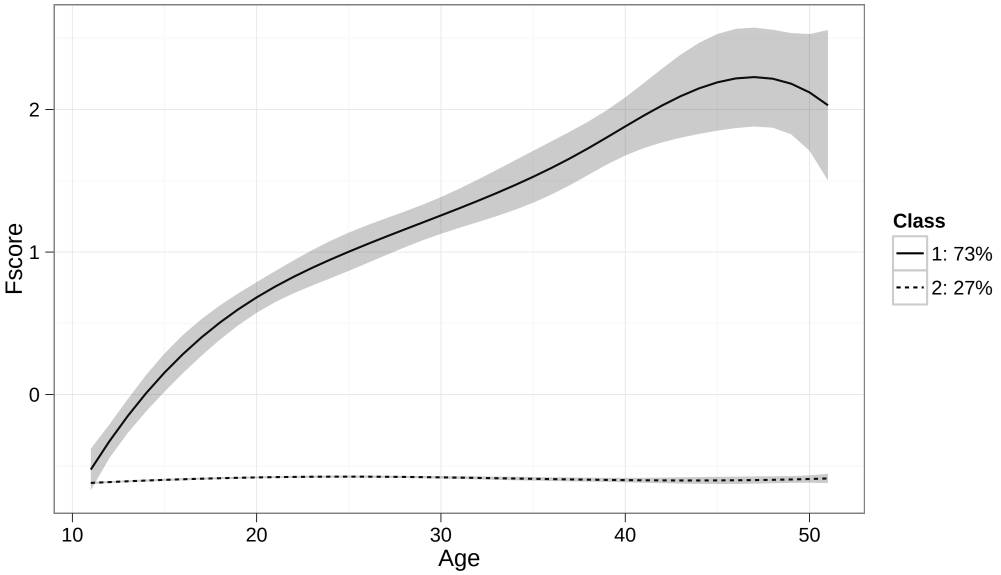


Figure 18: *Trajectories of factor three - two-class unconstrained model*

correlation between the two factors. The results in Table 13 seem to show strong positive associations between the active classes of Factor One and Factor Two, and strong negative associations between the active classes of the first two factors, and Factor Three.

5.e Evaluation of results from the factor model

The transformation of the frequencies of offending into a set of three factors led to three well-separated sets of trajectories. This separatedness is explained by the existence of inactive

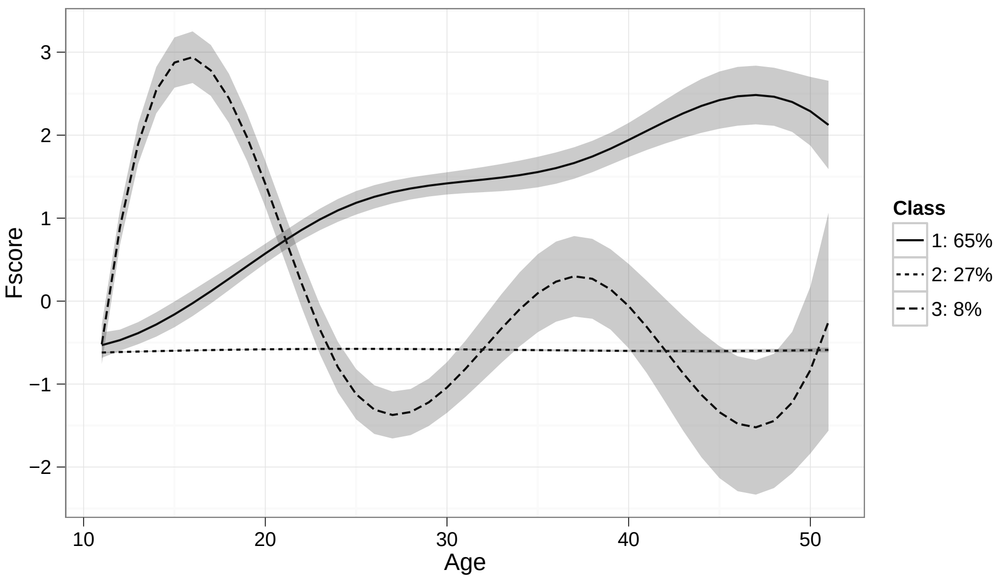


Figure 19: *Trajectories of factor three - three-class unconstrained model*

Table 12: Cross-classification of classes for Factors one-three

(a) Counts

Factor Two (class 1)

	Factor 3		
Factor 1	1	2	3
1	188	140	40
2	96	45	11
3	26	0	2

Factor Two (class 2)

	Factor 3		
Factor 1	1	2	3
1	84	15	20
2	79	9	1
3	44	2	5

(b) Proportions

Factor Two (class 1)

	Factor 3		
Factor 1	1	2	3
1	.233	.173	.050
2	.119	.056	.014
3	.032	.000	.002

Factor Two (class 2)

	Factor 3		
Factor 1	1	2	3
1	.104	.019	.025
2	.098	.011	.001
3	.055	.002	.006

classes of offenders in factors one and three. Indeed, the two-class models for factors one and three, consisting of only an active and an inactive class (not shown) produce perfect posterior classification of offenders. The factor models also produce fitted trajectories with, in general, smaller standard errors than the previous models. However, both the improved classification and the reduced imprecision are due to the fact that the factor scores do not take account of variation in the Poisson part of the factor model. The fitted trajectories are, in effect, the conditional means of conditional means, and the standard errors only take account of one level of variability.

Table 13: Results from a Poisson log-linear model on the cross-classified counts (two-way interactions only)

Interaction	Estimate	S.E.	Z	P value
F1(1) - F2(1)	1.43	0.26	5.43	< 0.0001
F1(2) - F2(1)	0.93	0.27	3.38	0.0007
F1(1) - F3(1)	-2.59	0.73	-3.55	0.0004
F1(2) - F3(1)	-2.10	0.74	-2.84	0.0046
F1(1) - F3(3)	-1.85	0.82	-2.25	0.0246
F1(2) - F3(3)	-2.51	0.87	-2.89	0.0039
F2(1) - F3(1)	-1.35	0.23	-5.82	< 0.0001
F2(1) - F3(3)	-1.20	0.32	-3.69	0.0002

The log-linear analysis of the cross-classified class assignements revealed a strong negative association between activity in factor three, and activity in the other two factors. If valid, this would suggest that those committing rape are less likely to commit child molestation and vice-versa, and would seem inconsistent with the observation that the largest sub-group of offenders are those who are active in all factors.

What the log-linear analysis really reveals is that it is impossible to estimate the association between rape and child molestation using a sample in which having committed one of these crimes is a necessary condition for sample membership. Generally, if the cross-classified table of probabilities of two events (say, commission of rape and child molestation) contains a cell for $P(\neg\texttt{rape}, \neg\texttt{child})$ that is structurally zero, then the log cross product ratio will be equal to $-\infty$ indicating perfect negative association. This will be the case if the occurrence of at least one of these events is a necessary condition for sample membership, even if the events are independent in the population. Although the log-linear analysis is not directly measuring the association between these two crimes, it is measuring the association between classes that are based upon factors, that are based upon frequencies of the two types of crime, and is therefore indirectly affected by the sampling restriction.

6 Conclusions

6.a Substantive research questions

The substantive research questions to be addressed by the study were categorised into the following three topics:

1. Bimodality of trajectories;

2. Association between child molestation and rape;

3. Association of non-contact sexual offending with serious offending.

The first question related to bimodality of trajectories concerned the existence, or not, of *individuals* with bimodal trajectories, as opposed to bimodality being an artefact of aggregation. The fitted curves of several GBTM classes[7] suggest that there are groups of people who have more than one peak of offending. The positions of the two peaks differ depending on the analysis. Furthermore, there is evidence from the factor model that the peak of rape, in the early 20s, is later than the peak for general offending, which occurs before the age of 20. This suggests that there may in fact be three peaks, with the adolescent peak of minor offending not well separated from the slightly later peak for rape.

We asked whether bimodality of trajectories could be explained by disaggregating crime types. The results of the factor model indicate that, although certain factors are associated with certain trajectories, there is nevertheless variation in trajectories after accounting for crime type. A surprising example of this is the group of "adolescent-limited" child molesters in the trajectory model for factor three (Figure 19). This echoes the small group of adolescent-limited sexual offenders that appeared in the dual trajectory analysis (Figure 14).

There is also strong evidence from the optimal matching (c.f. Figure 5) that those who commit rape in early life, particularly those who commit rape after the age of 20, are likely to commit crimes of child molestation in later life, as long as they are still active in later life. This evidence is corroborated by the factor analysis, particularly the cross-classified factor classes (c.f. Table 12), which show that the largest group of offenders is in class one or class two in factors one and two (active in rape and general offending), and class one in factor three (active in child molestation and non-contact sexual offending). There seems to be separation between the two within each time period, but a large overlap over the life course. The factor model, allowing the distinction between within-period, and within-individual association, is well suited to modeling this pattern. Specialisation in different crimes at different ages provides a possible explanation for the bimodality observed in the criminal careers of sexual offenders.

Non-contact sexual offending appears to be mostly associated with child molestation, although there is some evidence that it is associated with more specialised, less incidental, sexual offending in general. If valid, this finding might be useful in distinguishing between those whose sexual offending is part of a pattern of general offending, for example adolescent-limited sexual offenders, and sexual offenders with more persistent propensity to offend sexually.

[7] C.f. Figures 12, 16, 19.

Answering the second and third sets of research questions is limited, however, by sampling bias. It has been shown that it is not possible to quantify the extent of association between child molestation and rape over the life course, using a sample in which having committed one offence or the other is a necessary condition for sample membership. Similarly, it is not possible to quantify the extent to which non-contact sexual offending is associated with either of these serious crime types. This problem is not unique to this sample. Lussier (2010) and others have used datasets comprising only serious sexual offenders because they provide a level of detail that is difficult to find in datasets of general offenders. However, if these and other questions involving associations between types of crime are to be answered, samples of the wider population of offenders, or even including non-offenders, must be used.

6.b Evaluation of methods

6.b.i Optimal matching and non-probabilistic clustering

Optimal matching, far from being irrelevant, has been shown to be a useful complement to model-based trajectory modeling. The retrospective structure of the study partly obscures the extent to which hypotheses and research questions that motivated other analyses emerged from the use of optimal matching analysis as an exploratory method. Although limited by inability to take account of missingness such as periods of incarceration, and inability to make inferences from the results, it has the advantages that it is quick and easy to do, and no data aggregation is needed. However, ultimately we concur with Levine (2000) who argues that the idea of edit distance is not a convincing model for the underlying processes governing criminal careers, and that (overdispersed, zero-inflated) Poisson models are conceptually more valid. For this reason, optimal matching can only be useful as a complement to probabilistic models.

Conducting an analysis that relies on a completely different model and set of assumptions has the advantage that it can provide a sense-check for the main model. This is especially important in the case of GBTMs given their complexity and the dearth of diagnostics for conducting a thorough check of the fit of trajectories to observed counts.

6.b.ii Constrained and unconstrained group based trajectory models

The constrained and unconstrained trajectory models were found to both be useful in different contexts. When there is good correspondence between the group structure in one variable, and the group structure in another, the constrained model is a parsimonious model for common classes. However, by assuming a common group structure, the constrained model precludes the possibility of checking its validity. The unconstrained model is therefore a preferable starting point for analysis, since it allows the hypothesis of common group structure to be checked. We did not investigate the possibility of using the joint class membership probabilities from the unconstrained model to formally test hypotheses relating to common group structure, but we are aware that specific log-linear models can be formulated to test a variety of hypotheses about the structure of categorical associations. The major limitation of the unconstrained model is the fact that there are few software packages that can fit the full unconstrained model simultaneously.

6.b.iii The Poisson-log normal factor trajectory analysis

This study is the first, as far as we know, to use factor analysis to reduce the dimensionality of correlated counts, before subjecting the factor scores to a group based trajectory analysis. The concept for the model is similar to the "mixture of factor analysers" described by McLachlan and Peel (2000), except that the age co-efficients, rather than the factor loadings, were free to vary by class. Although still requiring refinement, the model fulfilled its intended purpose, which was to group crimes according to common trajectory paths. Crucially, the model allowed for the distinction between correlations at the level of five-year period, and associations between trajectories at the level of the individual. The emergent picture was of crimes types that were independent or even negatively associated in each period, but with a large degree of overlap over the criminal career.

One methodological question that was not dealt with in the study was whether the rotation of the factors should be done using an orthogonal or oblique method. Oblique methods are usually preferred in the social sciences because there is no reason why the factors (i.e. the latent constructs) should be assumed to be uncorrelated. However, when the factor scores are to be used in posterior analysis where one of the assumptions of the trajectory model is that the variables are independent within each component, allowing oblique rotation could invalidate this assumption. The trade-off is that orthogonal rotation algorithms are less likely to find simple structure.

Assessment of the goodness-of-fit of the fitted factor model was problematic because the usual fit statistics for linear factor models (proportion of variance explained, root-mean-squared error of approximation, covariance residuals, modification indices) could not be calculated, due to the lack of residual variance (apart from Poisson variance) in the model.

Originally when investigating the factor model it was envisaged that a sufficient statistic for the factors could be calculated on the scale of the crime counts, such that the sum of these statistics could be interpreted as the sum of the rates of the different types of crime. That is:

$$\sum_{l=1}^{q} f_{lt} = \sum_{g=1}^{p} \lambda_{gt}$$

The usefulness of this statistic would be in the interpretation as proportion of crime explained by the factor, and the factor scores would be on a meaningful scale. However, the log link meant that the measurement model was multiplicative for the λ_{gt}. It would be interesting to investigate whether an alternative gamma factor model with an identity link, as described by Wedel et al. (2003), would provide a solution to this problem.

6.b.iv Other methodological considerations

Other than the models themselves, a few other methods were used that were either innovative or at least uncommon in the context of criminal careers research.

The most important of these was the use of cubic splines in the linear model for the trajectories. The use of splines was central to answering the first set of research questions, regarding bimodality of trajectories.

Although Blokland et al. (2005) previously used this technique it does not seem to have gained in popularity since then, and the majority of researchers still employ quadratic or cubic polynomial terms in the specification of trajectories. Bushway et al. (2009) mentions splines, but chooses not to use them on the grounds that the addition of two additional terms adds too much complexity to the model. It is not clear why the maximum complexity of the linear model should be limited, as long as there are still enough degrees of freedom left to fit the model, and as long as all parameter estimates are significant[8]. Neither are splines more difficult to interpret than ordinary polynomials. Indeed, due to their local nature B-splines are easier to interpret than ordinary polynomials, because the sign and magnitude of each term can be interpreted as increasing or reducing the slope of the trajectory in that region of the observation period.

A criticism that applies equally to splines and other polynomials is that fitted curves can behave erratically in regions where data is sparse, such as near the end of the observation period. An adjustment that can be made to alleviate this problem is to constrain the degree of the spline to be linear after the last knot point, which is not possible with ordinary polynomials.

The use of the ICL-BIC, rather than an information criterion for the observed data likelihood, was also an uncommon choice for selecting from among group based trajectory models. As a fit statistic, the ICL-BIC tended to favour models with only one class in the single and dual trajectory analyses. This does not indicate a weakness of the measure. Rather it indicates that a zero-inflated negative binomial fit the data well, with little or no residual overdispersion. This was already hinted at before the models were fitted (c.f. Figure 8). If, instead of a negative binomial, a Poisson specification had been used (as is common in other studies), some of the groups identified by the model would have been artifacts of overdispersion, as was demonstrated in simulation studies by Skardhamar (2010). Nevertheless, in these instances we chose to analyse other models giving local minima of the ICL-BIC, in addition to the one-class solution. Perhaps this underlines the fact that fitting group based trajectory models will always have a subjective element, regardless of which selection criteria are used.

The zero-inflation adjustment that was used to account for unrecorded periods of incarceration was not uncommon in group based trajectory models. However, the method of its implementation, and its extension to deal with right-censoring, were unorthodox. Although the two-step process for calculating and then importing the (non-)zero-inflation weights into Latent Gold was reasonable, in hindsight we are not sure that our justification for extending the zero-inflation weights to deal with right-censoring was watertight. It is true that the missingness is non-ignorable, and would have caused bias in the trajectories. What is less clear is whether the adjustment we made alleviated the problem of bias, or just hid it. Perhaps the trajectories should have been presented as they were "warts and all" with the caveat that they were subject to bias. At any rate, we do not believe that the substantive conclusions would have been materially affected.

6.c Further work

Given the tentatively interesting substantive findings from the methods we have employed, it would be productive to repeat the analysis, using similar research questions and methods, but

[8] We have not presented significance of parameter estimates for GBTMs, but in general most of the spline terms were significant well beyond p=0.05.

on a larger sample with a less restrictive sampling scheme. This would hopefully enable the testing of associations between crime types, that was not possible using the current dataset.

Apart from this, there are a number of improvements that could be made to the method. The first of these would be to employ the full unconstrained model as described by Nagin and Tremblay (2001). This would ensure that there is no bias in the estimates of association between classes caused by classification errors in the modal assignment of observations to classes.

The factor model could possibly be improved in the ways described above, by replacing the log normal specification with a gamma specification for the factors themselves, with an identity link, and by constraining covariances between factors to zero.

The use of factor scores in posterior analysis leads to underestimates of the variance of estimates. This is because factor scores are estimates of unobserved random variables yet are treated as observed. One way around this would be to take a set of samples from the posterior predictive distribution of the factor scores (Asparouhov and Muthen, 2010), estimate the model repeatedly, and pool the estimates (Rubin, 2004).

This requirement, along with the ability to place a prior on the number of components, and the availability of posterior predictive checking as a way of thoroughly diagnosing model fit, points to the use of a Bayesian modeling paradigm for future development of the model. Bayesian methods lend themselves readily to latent variable models, since in a Bayesian framework the entire distribution of unknown quantities is estimated, rather than a point estimate of it.

References

Aitchison, J. and Ho, C. H. (1989). The Multivariate Poisson-log normal Distribution. *Biometrika*, 76(4):643–653.

Asparouhov, T. and Muthen, B. (2010). Plausible Values for Latent Variables using MPlus. Unpublished. Accessed at: `http://www.statmodel.com/download/Plausible.pdf`.

Bartholomew, D. J. and Knott, M. (1999). *Latent Variable Models and Factor Analysis.* Kendall's Library of Statistics 7.

Bartolucci, F., Pennoni, F., and Francis, B. (2007). A latent Markov model for detecting patterns of criminal activity. *Journal of the Royal Statistical Society Series A - Statistics in Society*, 170(Part 1):115–132.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725.

Blokland, A., Nagin, D., and Nieuwbeerta, P. (2005). Life Span Offending Trajectories of a Dutch Conviction Cohort. *Criminology*, 43(4):919–954.

Brame, B., Nagin, D., and Tremblay, R. (2001). Developmental trajectories of physical aggression from school entry to late adolescence. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 42(4):503–512.

Bushway, S. D., Sweeten, G., and Nieuwbeerta, P. (2009). Measuring Long Term Individual Trajectories of Offending Using Multiple Methods. *Journal of Quantitative Criminology*, 25(3):259–286.

Eggleston, E., Laub, J., and Sampson, R. (2004). Methodological sensitivities to latent class analysis of long-term criminal trajectories. *Journal of Quantitative Criminology*, 20(1):1–26.

Farrington, D. P. (1986). Age and Crime. *Crime and Justice*, 7:189–250.

Gabadinho, A., Ritschard, G., Muller, N., and Studer, M. (2011). Analysing and Visualizing State Sequences in R with TraMineR. *Journal of Statistical Software*, 40(4):1–37.

Gottfredson, M. and Hirschi, T. (1990). *A General Theory of Crime*. Sociology / Stanford. Stanford University Press.

Hanson, R. (2002). Recidivism and age - Follow-up data from 4,673 sexual offenders. *Journal of Interpersonal Violence*, 17(10):1046–1062.

Harris, D., Smallbone, S., Dennison, S., and Knight, R. (2009). Specialization and Versatility in Sexual Offenders Referred for Civil Commitment. *Journal of Criminal Justice*, 37:37–44.

Haviland, A., Jones, B., and Nagin, D. (2011). Group-Based Trajectory modeling extended to account for nonrandom subject attrition. *Sociological Methods and Research*, 40:367–390.

Hinde, J. and Demetrio, C. (2010). *Overdispersion: Models and Estimation*. Chapman and Hall/CRC Interdisciplinary Statistics Series. Taylor & Francis.

Hirschi, T. and Gottfredson, M. (1983). Age and the Explanation of Crime. *American Journal of Sociology*, 89(3):552–584.

Keele, L. (2008). *Semi-parametric Regression for the Social Sciences*. Wiley.

Levine, J. (2000). But what have you done for us lately? Commentary on Abbott and Tsay. *Sociological Methods & Research*, 29(1):34–40.

Lussier, P. (2010). Criminal Trajectories of Adult Sex Offenders and the Age Effect: Examining the Dynamic Aspect of Offending in Adulthood. *International Criminal Justice Review*, 20(2):147–168.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

McVicar, D. and Anyadike-Danes, M. (2010). Does Optimal Matching Really Give Us Anything Extra for the Analysis of Careers Data? An Application to British Criminal Careers. Unpublished. Accessed at: `http://www.qub-efrg.com/fs/doc/working-papers/mcvicardanes2010crime.pdf`.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial-behaviour - a developmental taxonomy. *Psychological Review*, 100(4):674–701.

Muthen, L. K. and Muthen, B. O. (1998-2011). *MPlus User's Guide*. Los Angeles, CA, sixth edition.

Nagin, D. (2005). *Group-Based Modeling of Development*. Harvard University Press.

Nagin, D. and Land, K. (1993). Age, Criminal Careers, And Population Heterogeneity - Specification And Estimation Of A Nonparametric, Mixed Poisson Model. *Criminology*, 31(3):327–362.

Nagin, D. and Tremblay, R. (2001). Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological Methods*, 6(1):18–34.

Pedneault, A., Harris, D. A., and Knight, R. A. (2012). Toward a typology of sexual burglary: Latent class findings. *Journal of Criminal Justice*, 40(4):278 – 284.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2.15.1 edition. ISBN 3-900051-07-0.

Rubin, D. (2004). *Multiple Imputation for Nonresponse in Surveys*. Wiley Classics Library. Wiley-Interscience.

Skardhamar, T. (2010). Distinguishing Facts and Artifacts in Group-Based Modeling. *Criminology*, 48(1):295–320.

Soothill, K. and Francis, B. (1999). Reviewing the Pantheon of Sexual Offending. *Amicus Curiae: Journal of the Society for Advanced Legal Studies*, 17:4 – 8.

Soothill, K., Francis, B., Sanderson, B., and Ackerley, E. (2000). Sex Offenders: Specialists, Generalists - or Both? *The British Journal of Criminology*, 40(1):56–67.

Studer, M., Muller, N. S., Ritschard, G., and Gabadinho, A. (2010). Classer, discriminer et visualiser des sequences d'evenements. (in press). Accessed at: `http://mephisto.unige.ch/pub/publications/gr/Matthias-EGC10-paper37.pdf`.

Vermunt, J. and Magidson, J. (2005). *Technical Guide for Latent GOLD 4.0: Basic and Advanced*. Statistical Innovations Inc., Belmont Massachusetts.

Weakliem, D. L. and Wright, B. R. E. (2009). Robustness of Group-Based Models for Longitudinal Count Data. *Sociological Methods & Research*, 38(1):147–170.

Wedel, M., Bockenholt, U., and Kamakura, W. A. (2003). Factor models for multivariate count data. *Journal of Multivariate Analysis*, 87:356 – 369.