School of Systems Engineering

Using Topic Models to Detect Behaviour Patterns for Healthcare Monitoring

Thesis submitted for the degree of Doctor of Philosophy

Ruth Jemma White

Supervisors:
Professor William Holderbaum
Professor William Harwin

September 2017

To those lost along the way,

Jean Leslie Evans
26th April 1928 - 12th February 2015

Harneet Arora
15th September 1987 - 30th December 2016

Declaration of original authorship

Declaration: I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Ruth Jemma White

Abstract

Healthcare systems worldwide are facing growing demands on their resources due to an ageing population and increase in prevalence of chronic diseases. Innovative residential healthcare monitoring systems, using a variety of sensors are being developed to help address these needs. Interpreting the vast wealth of data generated is key to fully exploiting the benefits offered by a monitoring system. This thesis presents the application of topic models, a machine learning algorithm, to detect behaviour patterns in different types of data produced by a monitoring system. Latent Dirichlet Allocation was applied to real world activity data with corresponding ground truth labels of daily routines. The results from an existing dataset and a novel dataset collected using a custom mobile phone app, demonstrated that the patterns found are equivalent of routines. Long term monitoring can identify changes that could indicate an alteration in health status. Dynamic topic models were applied to simulated long term activity datasets to detect changes in the structure of daily routines. It was shown that the changes occurring in the simulated data can successfully be detected. This result suggests potential for dynamic topic models to identify changes in routines that could aid early diagnosis of chronic diseases. Furthermore, chronic conditions, such as diabetes and obesity, are related to quality of diet. Current research findings on the association between eating behaviours, especially snacking, and the impact on diet quality and health are often conflicting. One problem is the lack of consistent definitions for different types of eating event. The novel application of Latent Dirichlet Allocation to three nutrition datasets is described. The results demonstrated that combinations of food groups representative of eating event types can be detected. Moreover, labels assigned to these combinations showed good agreement with alternative methods for labelling eating event types.

Acknowledgements

I would like to acknowledge the funding for this PhD by the University of Reading Research Endowment Trust Fund (RETF), without this scholarship it would not have been possible to conduct this research. In addition, this PhD was associated with the SPHERE IRC project funded by the UK Engineering and Physical Sciences Research Council (EPSRC), Grant EP/K031910/1.

I would like to take this opportunity to offer my heartfelt thanks to my PhD supervisors, Professor William Harwin and Professor William Holderbaum, for all of their support and guidance throughout my research. A special mention goes to William Harwin for believing in my potential and encouraging me along the academic path since the second year of my undergraduate degree. Thank you for all your inspirational and interesting chats. Many thanks also go to my nutritionist collaborators, Dr Laura Johnson from the University of Bristol and Dr Eileen Gibney from University College Dublin. I appreciate the time you took to share your knowledge, ideas and data with me. It is always refreshing to see things from an alternative point of view and work across disciplines to discover innovative ways of solving problems.

I would like to thank all the members of the SPHERE project team, past and present for your support, friendship, collaboration and intellectual stimulation. It was great to be part of a big project family, in particular the chance to share ideas in workshops and on away days, especially with those from other disciplines. Special thanks to everyone in the Reading office for keeping me going over the last few years, it wouldn't have been the same without you. I would also like to thank everyone in the School of Systems Engineering (sadly no more!), especially Dr Hwang, Dr Janko, Dr Cave-Ayland, Mr Salmon, Dr Zoulias, Ms Fannon, Mr Laird, Dr Tse, Ms Cheung and Mr Gould, for all your friendship, inspiration and pearls of wisdom over the years, since starting my first degree. I would never have got this far without you all!

The academic road can be a rocky one and the summit sometimes dips out of sight; many thanks go to all my mountaineering and climbing friends. Our journeys along real rocky paths to summits with beautiful views (when not covered in cloud!) provide much needed respite from a computer screen and blow the cobwebs away to make space for fresh ideas and solutions. Thank you for listening and sharing these moments with me. An extra thank you to Ania, Pete (and Merlin) for sharing your first house with me and giving me a wonderful home to live in. Many thanks also go to the other rocks in my life, both past and present, for all the much needed fun, laughter and long chats. You know who you are, thanks!!

You don't easily get far in life without the dedication and support of a family, in all shapes and sizes, financially, emotionally and much more! Over the course of this PhD I've had the great joy of two new additions in the form of my gorgeous niece and nephew, such breaths of fresh air. I love to watch you grow and learn. My grandpa has challenged and inspired me with his stories and questions during many visits in hospitals and nursing homes. My mum and dad are *always* there, no matter what or when, even if it means a road trip down the Florida Keys after presenting at a conference ;)

Last, but by no means least, I'd like to thank my fiancé, Vinny. Who knows, if it wasn't for my PhD I might not have taken an interest in your wearable device and we wouldn't be where we are today! You might not have been there since the start but your input has been invaluable. Thank you for all your love, patience and advice, especially for keeping me going through the write up. This is just the start, I can't wait for the rest of our adventures together :)

The latex template used is based on a style by C. Dean Barnes and Chris Duggan of Memorial University, Newfoundland, Canada.

Table of Contents

List of Tables

List of Figures

Mathematical Notation

Within the machine learning community a variety of different mathematical notations are often used to represent the same concept. For clarification an explanation of the notation used in this thesis is given.

$\vec{x} = (x_1, x_2, \ldots, x_K)$ denotes a vector of K elements and is often shortened to $\vec{x}$ to increase readability. This notation is used to denote a discrete multinomial variable that can take one of K possible mutually exclusive states. In this case, one element x_k will equal 1 and all other elements will equal 0.

$x_{1:N}$ denotes a range of random variables from x_1 to x_N

$\vec{x}_{1:N}$ denotes a range of multinomial random variables from $\vec{x}_1$ to $\vec{x}_N$

$p(x)$ denotes the probability distribution over the random variable x. This is the marginal probability.

$p(x, y)$ denotes the joint probability of x and y

$p(x|y)$ denotes the conditional probability of x given y

$\mathbb{E}_x[f(x, y)]$ denotes the expectation of a function $f(x, y)$ with respect to a random variable x

List of Abbreviations

AAL	Ambient Assisted Living
ADL	Activities of Daily Living
BLE	Bluetooth Low Energy
CASAS	Center for Advanced Studies in Adaptive Systems
COPD	Chronic Obstructive Pulmonary Disease
CRP	Chinese Restaurant Process
CRS	Circadian Rhythm Score
CTM	Correlated Topic Model
ddCRP	distance dependent Chinese Restaurant Process
dHDP	Dynamic Hierarchical Dirchlet Processes
DP	Dirichlet Process
DTM	Dynamic Topic Model
ECG	Electrocardiogram
ELBO	Evidence Lower Bound
EM	Expectation Maximisation
EPSRC	Engineering and Physical Sciences Research Council
GDP	Gross Domestic Product
GSM	Global System for Mobile Communications
HDP	Hierarchical Dirchlet Processes
HMM	Hidden Markov Model
IC	Intensity Categories
IUNA	Irish Universities Nutrition Alliance
JDL	Joint Directors of Laboratories
KL	Kullback-Leibler
LDA	Latent Dirichlet Allocation
MCMC	Markov Chain Monte Carlo
MIT	Massachusetts Institute of Technology
NANS	National Adult Nutrition Survey
NDNS 2000	National Diet and Nutrition Survey 2000
NDNS RP	National Diet and Nutrition Survey Rolling Programme

NHS	National Health Service
PA	Physical Activity
PCA	Principle Components Analysis
PIR	Passive Infra-Red
RGB-D	Red, Green, Blue and Depth
SPHERE IRC	Sensor Platform for Healthcare in a Residential Environment Interdisciplinary Research Collaboration
TEI	Total Energy Intake

Chapter 1

Introduction

There is a demand from the public, in countries across the globe, for affordable access to high quality health care [1]. This goal poses many unprecedented challenges that must be tackled. Several new approaches and initiatives are being considered in an effort to address these needs, including the role of technology in healthcare systems [2,3]. In particular, empowering people to monitor their health and well-being at home is gaining increasing interest from interdisciplinary research communities [4]. Residential healthcare monitoring systems can generate a vast amount of data from a variety of sensors. Interpreting this data is key to fully exploiting the benefits offered by a monitoring system. Data fusion and machine learning techniques can be employed to detect patterns and provide useful information to the end users.

This thesis explores high level data fusion for healthcare monitoring systems. The main focus is on the use of the machine learning algorithm, topic models, originally developed by the text processing community. The application of topic models to detect behaviour patterns from different types of data produced by a residential monitoring system is presented and the results are analysed and discussed.

1.1 Motivation

The UK's National Health Service (NHS) is facing ever growing demands on its resources in order to continue to be able to provide world class health care. One key factor is the demographic change that is occurring. It is predicted that approximately 26% of the UK population will be aged 65 or over by 2061, compared with 17% in 2012, leading to a rise in the proportion of gross domestic product (GDP) that is spent on health and social care [5]. In particular, it is projected that by 2065 age-related spending on health and long-term care will increase to 10.2% of GDP from an estimated 8.4% in 2015 [6]. An expansion in morbidity is another contributing factor. Although there may be an increase in the average healthy life expectancy it is likely that the absolute amount of time spent in poor health and associated costs will still increase overall [5].

Another critical factor is the increase in prevalence of chronic diseases and conditions, such as depression, diabetes and heart failure. For example, approximately 4% of the UK population have diabetes and the treatment for this requires about 10% of the NHS budget [7]. Although this increase is in part related to the demographic change, it is also due to lifestyle and environmental changes, such as diet and nutrition intake [8]. In fact, for current generations, chronic diseases and the associated disabilities could well have a larger impact on healthcare systems than age-specific disabilities [9]. This is an issue that resonates worldwide, as highlighted by the Director General of the World Health Organisation in the 2010 World Health Report [1]. Over the past 25 years, the global burden of neurological disorders, including dementia, Parkinson's disease and epilepsy, has substantially increased. This group is the leading cause of disability, with stroke being the largest contributor [10].

Many chronic diseases are related to physical activity (PA) levels and these can be a relevant indicator of a person's health, particularly for COPD patients [11, 12]. Low levels of PA and corresponding energy expenditure are associated with greater risk of hospital admission and mortality. Rehabilitation programs for pulmonary diseases aim to increase and maintain PA levels over time, including when a patient returns to their home after discharge [12]. Research has also shown that PA levels are inversely related to cardio-metabolic risk factors in youth [13]. Furthermore, PA is important for retaining the ability to conduct activities of daily living (ADL), however the recommended guideline levels of PA are often not met by older adults [14]. Monitoring a person's ADL, such as eating, showering and sleeping, can be used as a measure of functional health [3, 15].

In addition to PA, chronic diseases are also often related to diet, such as diabetes and obesity [16]. To help tackle such problems a variety of methods are employed, including dietary interventions focussed on achieving a nutritionally well-balanced diet. However, findings from nutrition research are often conflicting with regards to the impact of snacking on diet quality and health. One problem with interpreting the results of relevant studies is the lack of a clear, unified definition of snacking [16–18]. Various approaches have been used throughout the literature. Some studies focus primarily on snacking, others consider what constitutes a meal and some researchers use a more neutral definition of eating events or occasions, to include any ingestion of food or drink. As well as the wide range of methods used in studies, the definitions themselves can often be based on cultural norms and hence be biased by the experiences of the researcher [18]. This lack of agreement restricts the dissemination of information to the public regarding the impact of snacking on health [16].

1.2 Long term healthcare monitoring in a residential environment

Long term monitoring of chronic conditions can aid early diagnosis and enable them to be managed more effectively [3, 11, 19, 20]. It is important to assess patients in a clinical setting, however there are often severe time constraints and limited resources. Therefore observations of a patient may only give a snapshot of their condition instead of the full picture [21,22]. There is interest in developing monitoring systems that can be used in a residential environment to help alleviate the strain on healthcare systems across the globe [19, 23] and enable patients to remain independent in their own home for longer, thus improving quality of life [24]. Moreover, a remote monitoring system could be used to assist falls prevention and detect emergency events and hence provide a rapid response [3]. This avoids scenarios where people are left lying on the floor for hours or even days before they are found and hence reduces the risk of further complications that result from not receiving care in a timely manner, such as pneumonia [25].

The development of monitoring systems for use in a residential environment are gaining interest across several multi-disciplinary research communities [3,4,14,23–26]. In general, monitoring systems are built up of a network of sensors, which can collect a range of data about a person in their home. This can include wearable sensors, that could be used to detect movement or even vital signs, such as heart rate [3]. Environmental sensors to monitor temperature, humidity, electricity usage, water usage and movement can also be installed throughout the house [27]. Additionally, RGB-D (visual image and depth) cameras may be used to detect a person's motion and activities [28]. The sensors in the network communicate with a central node via wireless protocols, such as Bluetooth Low Energy (BLE) or Zigbee, or a wired connection. The data can be fused and the results displayed on a user interface and transmitted securely over the internet. This allows data to be shared with key stakeholders, such as the individual, their family members, carers or clinicians [23].

The SPHERE (a Sensor Platform for HEalthcare in a Residential Environment) IRC (Interdisciplinary Research Collaboration) is an EPSRC (Engineering and Physical Sciences Research Council) funded project, with which this PhD is associated. The aim of the project is summed up in the project mission statement:

> "The vision of SPHERE is not to develop fundamentally-new sensor technology specifically for individual disease conditions but rather to impact a range of healthcare needs simultaneously by employing data-fusion and pattern-recognition from a common platform of largely non-medical/environmental networked sensors in a home environment" [29]

The SPHERE IRC project commenced in October 2013 and had initial funding for a duration of 5 years. The sensor platform developed was deployed in the homes of members of the public in year 4 of the project. Additional funding has recently been secured to continue the project until 2021.

The nature of the SPHERE IRC means that clinicians, engineers, computer scientists, designers and many others have been collaborating on this challenge from day one in order to develop a user friendly and effective system. In particular, the project is exploring the use of energy harvesting techniques and low power wireless technologies to develop a wearable device that does not require batteries to be changed or plugged in to charge. This makes the system easier to use in a long term monitoring situation, particularly if the user is not able to easily interact with technology [29].

The project is using three main types of sensors: environmental sensors; video and depth cameras and wearable sensors, such as accelerometers [29]. Each sensing modality has its strengths and weaknesses but by combining the data from all of them the best possible picture of a person's activities within the house can be created [30]. For example, there are privacy concerns related to the use of cameras in some locations such as the bathroom or bedroom but these areas can be monitored through wearable and environmental sensors that are less obtrusive. On the other hand, an environmental sensor would not be able to detect what a person is eating but a camera that is correctly aligned would be able to distinguish between different types of food [31].

1.3 Research aims and challenges

The aim of the research conducted for this thesis is to detect behaviour patterns in data collected by a residential healthcare monitoring system, such as being developed by the SPHERE IRC, in order to obtain information of a greater quality than the raw sensor data. There are many research challenges in this area, including: user acceptance, privacy concerns, data storage and processing, establishing ground truth references, annotation, power requirements and cost to name a few. This section highlights in more detail the research challenges and the corresponding research questions identified as the focus for this thesis.

Detection of daily routines for healthcare monitoring. A residential healthcare monitoring system can generate a vast wealth of data from a variety of sources. This complex data is difficult to interpret and is not of direct use to clinicians, caregivers or individuals being monitored. Recognising activities from the data reduces the quantity and complexity to some degree but generally the discovered activities are still at a low level of abstraction. Determining how these activities are structured together in higher level routines can facilitate efficient and effective extraction of relevant information from the data. This knowledge can be used to inform decision making processes related to health and well-being. Existing work has demonstrated the successful use of topic models to detect daily routines from activity data, based on seven days of real world data collected for one person wearing two accelerometers [32]. To determine whether topic models are extendible to other real world datasets with different activities and routines is a current research challenge. This can be formulated as the first of three research questions:
Can patterns in a person's daily activities that are representative of routines be detected for a novel, real world activity dataset?

Identification of changes in routines over time. One key benefit of monitoring a person in their own home is that a more accurate picture of their state of well-being and health can be acquired over an extended period of time. This is in comparison to the information that can be gathered from a snapshot situation, such as an outpatient appointment at a hospital or a visit to a GP. Current research in long term monitoring and detection of changes in behaviour patterns has focussed on variation in activities or intensity levels. This can result in too much information to process efficiently if there are a lot of activities or the loss of important details if the summary is at a high level. Changes over time in the structure of daily routines has not previously been investigated. The benefit of this approach is that it enables both a high level summary whilst still incorporating useful information from the more detailed data. This challenge can be formulated as the second research question:
Can changes in the structure of daily routines over time be detected?

Understanding how foods are combined in eating events. The data generated by a residential healthcare monitoring system does not only relate to activities of daily living. Other types of data can also be collected, such as nutrition data in the form of dietary intake. Diet and eating behaviours can be linked with health outcomes and the management of chronic diseases. Presently there are myriad definitions of types of eating events, such as a meal or snack. This poses a significant barrier to comparing results across different studies and establishing clear links between eating behaviours and the impact on diet quality and health. The existing definitions are limited and often influenced by language and cultural norms. Understanding how

people actually combine foods in different types of eating events based on real data could facilitate analysis of how eating behaviours and health outcomes are related. This challenge can be formulated as the final research question:
Can the combinations of foods consumed together in different types of eating events be automatically detected?

1.4 Contributions to knowledge

The contributions to knowledge for this thesis follow on from the research challenges identified and the formulation of associated research questions. These contributions are:

- ▷ Validation of application of topic models for detecting patterns in activities reflective of daily routines for the UbiComp 08 dataset first published by Huynh et al. [32].
- ▷ Development of a novel mobile phone application for logging activities of daily living and associated routine labels. Knowledge gained through this was shared with the wider SPHERE project to help develop a similar application for annotating datasets collected using the SPHERE sensor platform.
- ▷ Collection of novel activity and daily routine dataset for two participants with 16 days of data each.
- ▷ Application of topic models to detect patterns in activities reflective of daily routines for a novel dataset.
- ▷ Method for implementation of dynamic topic models to activity data for detecting changes in routines over time.
- ▷ Evaluation of impact of activity data properties on detecting changes in routines over time using dynamic topic models applied to a simulated dataset with known variations.
- ▷ Application of dynamic topic models to simulated activity dataset to detect unknown changes in routines over time.
- ▷ Method for applying topic models to nutrition data to detect how food groups are combined together in different types of eating events.
- ▷ Application of topic models to food diary data in the National Diet and Nutrition Survey Rolling Programme dataset and analysis of the resulting food group combinations detected.

- ▷ Evaluation of the impact of dataset properties on the combinations of food groups found by topic models.
- ▷ Utilisation of the combinations of food groups consumed together in an eating event to visualise underlying patterns in the dataset.
- ▷ Validation of the application of topic models to food diary data using the National Diet and Nutrition Survey 2000 dataset, and comparing the resulting food group combinations to those found using a manual rule-based approach.
- ▷ Validation of the application of topic models to food diary data using the Irish National Adult Nutrition Survey, and comparing the resulting food group combinations with participant defined labels.

Part of this thesis, relating to the work presented in chapter 6, has been published in the peer reviewed International Conference on Machine Learning and Applications [33]. Journal publications are planned for the work detailed in chapters 5 and 7. The literature review undertaken as part of this thesis, chapter 3, contributed to a review paper published in the Medical Engineering and Physics Journal [34]. In addition, the author contributed to the set-up of initial experiments to investigate the minimum sensor separation of two accelerometers to achieve a specified signal to noise ratio in the application of recognising human motions, published in IEEE Sensors Journal [35].

1.5 Thesis structure

An outline of the structure of this thesis is given here to facilitate navigation of the document.

Chapter 2 provides the reader with background information on probability theory and machine learning, with a strong focus on topic models.

Chapter 3 reviews the literature in the field of data fusion techniques for applications related to healthcare monitoring in a residential environment. This includes sections on detecting behaviour patterns in activity data synonymous with daily routines and their changes over time, as well as behaviour patterns relating to nutrition data and definitions of eating event types.

Details of the implementation of topic models to detect daily routines in activity data for an existing and a novel dataset and analysis of the results are given in chapter 4. The development of a mobile phone application to log activities of daily living and corresponding routines and its implementation to collect a novel dataset is described.

Chapter 5 presents the application of dynamic topic models to activity data to detect changes in routines over time in terms of the probability of constituent activities. The method is developed for simulated datasets with known changes and validated using simulated datasets with variations that are unknown a priori.

The development of a method to apply topic models to nutrition data is outlined in chapter 6 using the National Diet and Nutrition Survey Rolling Programme dataset. The results revealing the structure of how foods are combined in different types of eating events are presented and discussed.

The utilisation of the eating event types discovered using topic models to investigate patterns of eating behaviour at the individual and population levels are explored in chapter 7. The application of topic models for discovering eating event types are validated for two further datasets. The results are compared with existing methods for labelling eating event types.

Finally, conclusions of the research presented are drawn in chapter 8 and avenues for future work are highlighted.

Chapter 2

An Introduction to Topic Models

Extracting useful information from large collections of data, such as those generated by a residential healthcare monitoring system, is difficult. Unsupervised machine learning algorithms provide powerful tools for finding patterns and hidden structure in data [36]. Moreover, probabilistic models enable uncertainty in the data and underlying patterns to be modelled [37]. In particular, topic modelling algorithms, originally developed to discover hidden themes in a large corpus of text documents, have gained increasing popularity and have been used to find patterns in a variety of data sources. The resulting distributions discovered in the data can be used for analysing, summarising, searching, clustering and exploring the original large, complex dataset [36]. This chapter introduces the background theory on machine learning, with a focus on topic models, that underpins the remainder of the work presented in this thesis. These concepts are described here to facilitate the analysis of the literature reviewed in chapter 3.

2.1 Background probability theory and concepts

This section outlines the key concepts involved in the understanding of topic models and how they are applied. The Bayesian view of probability theory is contrasted against the frequentist view and the underlying Dirichlet distribution used in topic models is introduced. An invaluable tool, probabilistic graphical models, are introduced as these will be used to represent topic models. Finally, a brief overview of the concept behind topic models and the associated modelling assumptions are described.

2.1.1 Probability theory and Dirichlet distribution

Probability theory is vital to machine learning as it provides a consistent framework for dealing with uncertainty. There are two main views of probability. Firstly, the frequentist view, where probabilities are in terms of the frequencies of random, repeatable events. Secondly, the more general Bayesian view that probabilities provide a quantification of uncertainty. Generally it is more useful to adopt the Bayesian interpretation of probability for machine learning problems. This approach allows new evidence to be used to revise a quantification of uncertainty. The advantages and disadvantages of each viewpoint are given below [38]:

Bayesian

+ Inclusion of prior knowledge occurs naturally
+ Deterministic approximation schemes allow use of techniques in large-scale applications
– Often criticised that selection of prior is subjective
– Computationally expensive

Frequentist

+ Easy to compare models
– Can lead to extreme conclusions

Probabilities can follow different distributions over a random variable, such as the binomial distribution for discrete variables and the Gaussian distribution for continuous variables. The exact nature of these distributions are controlled by parameters, for example the mean and variance for a Gaussian, hence they are examples of parametric distributions. Taking a Bayesian approach to determining suitable values for the parameters for a particular dataset, a prior distribution can be placed over the parameters. New evidence, in the form of observed data, can be incorporated by applying Bayes' theorem to compute the corresponding posterior distribution. This Bayesian analysis can be greatly simplified by using conjugate priors, meaning that the posterior distribution has the same functional form as the prior [38]. In particular, the Dirichlet distribution has the same functional form as the multinomial distribution and is therefore the conjugate prior.

More specifically, consider a multinomial discrete variable as a vector of K mutually exclusive states $\vec{x} = (x_1, x_2, \ldots, x_K)$, such that only one state, x_k, can equal 1 and all other states are 0. For example, if a variable can take $K = 4$ states, $\vec{x} = (0, 1, 0, 0)$ represents the observation of the variable in the state where $x_2 = 1$. Let the parameter μ_k be the probability that $x_k = 1$, hence the vector of parameters can be written as $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)$ and satisfies the conditions $\mu_k \geqslant 0$ and $\sum_{k=1}^{K} \mu_k = 1$. Then for N independent observations, $\vec{x}_{1:N}$, the number of observations of each state, $x_k = 1$, can be represented by $m_k = \sum_{n=1}^{N} x_{n,k}$. The multinomial distribution, equation 2.1, is the joint distribution of the quantities m_k, for k from 1 to K, conditioned on the parameters $\vec{\mu}$ and the total number of observations N [38].

$$Mult(m_1, m_2, \ldots, m_K|\vec{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k} \tag{2.1}$$

where $\binom{N}{m_1 m_2 \ldots m_K} = \frac{N!}{m_1! m_2! \ldots m_K!}$.

Furthermore, the conjugate prior of the multinomial distribution is the Dirichlet distribution over the parameters $\vec{\mu}$, which is confined to the open simplex of dimensionality $K-1$, due to the summation constraint. The Dirichlet distribution, equation 2.2, has hyperparameters $\vec{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_K)$ which control the form of the distribution [38].

$$Dir(\vec{\mu}|\vec{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \tag{2.2}$$

where, Γ is the gamma function defined as $\Gamma(x) \equiv \int_0^\infty u^{x-1} e^{-u} du$.

For a Dirichlet distribution over 3 variables, with all $\alpha_k = 1$, the $K-1$ simplex can be visualised as a flat triangle, as shown in figure 2.1, where all points on the simplex have equal probability of being sampled. The figure shows graphs of three potential distributions that could be sampled from the Dirichlet and their corresponding locations on the simplex. It should be noted that the distributions corresponding to the edge of the simplex cannot actually be achieved but can be arbitrarily close, that is, no component of a distribution drawn from a Dirichlet will ever be zero. This is due to the fact that the support of the Dirichlet is the open simplex [39,40]. For $\alpha_k < 1$

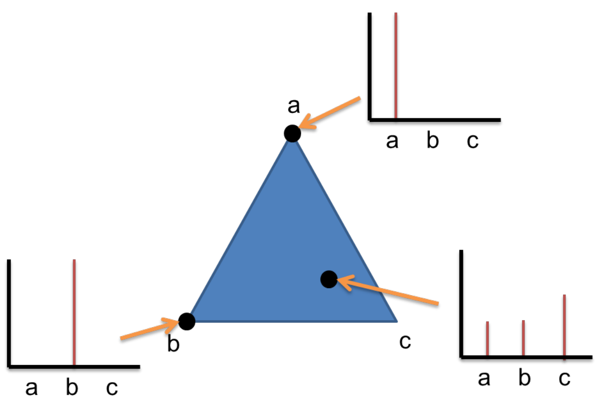
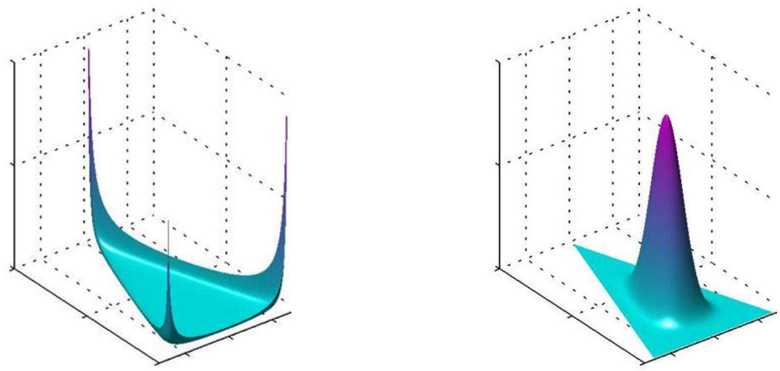


Figure 2.1: Sample distributions from a Dirichlet with all $\alpha_k = 1$



(a) $\alpha_k = 0.1$ (b) $\alpha_k = 10$

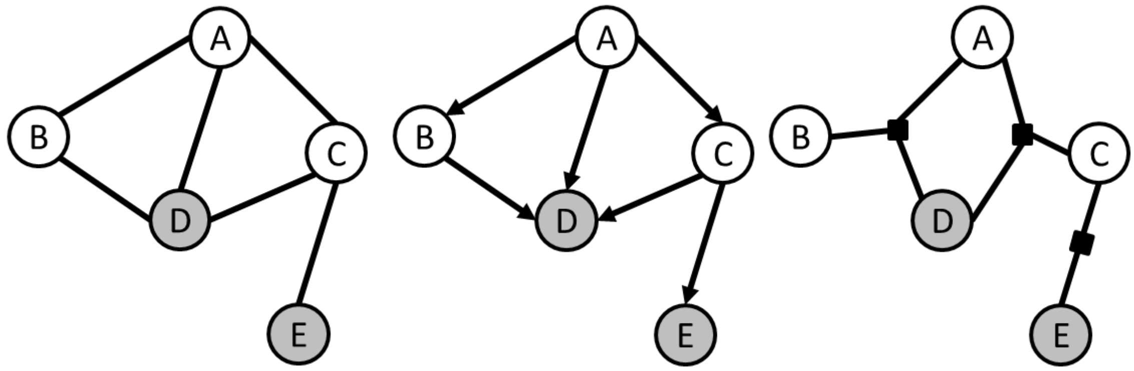
Figure 2.2: Dirichlet distributions with different values for the α hyperparameter Taken from [38] under permission granted for non-commercial use.

the Dirichlet becomes much more sparse, as shown in figure 2.2a and for $\alpha_k > 1$ the Dirichlet is peakier around the expectation of the distribution, as shown in figure 2.2b [41].

The Dirichlet Process (DP) provides a distribution on distributions over an arbitrary space, $G \sim DP(\alpha, G_0)$ where G is a random distribution over some space, α is the precision parameter, a positive scalar and G_0 is the known base distribution over the same space as G. There are two important properties of the Dirichlet Process. Firstly, draws from the DP are discrete, with positive probability mass at atoms generated independently from G_0. An atom is a measurable set which has positive measure and contains no set of smaller positive measure [42]. A low value of α means that there are a few dominant atoms, whilst a high value gives a discrete distribution more similar to G_0. Secondly, the clustering property means that draws from G, if drawn itself from a DP, can be partitioned according to atoms they share [36].

2.1.2 Probabilistic graphical models

Probabilistic graphical models provide a diagrammatic representation of probability distributions and are a very effective way of visualising and designing the structure of probabilistic models, rather than just using algebraic manipulations. They provide key insights into properties of a model, simply through inspection of the graph and can help simplify inference problems [38]. Graphical models lie at the boundary of graph theory and probability theory; providing a formalism for understanding and generalising the development and application of statistical models [43]. There are three main types of probabilistic graphical models: undirected graphs, directed graphs and factor graphs, as shown in figure 2.3. All of these use nodes, representing random variables, connected by edges, representing statistical dependencies between variables.



(a) Undirected graph (b) Directed graph (c) Factor graph

Figure 2.3: Examples of the three main types of graphical models

Shaded nodes are observed variables and unshaded nodes are latent (hidden) variables [44]. Each type of graphical model is better suited to expressing different aspects of the model and it is possible to convert between them if required [38].

In this thesis, the focus is on directed acyclic graphs, also known as Bayesian networks or belief networks. The arrows in these graphs represent causal relationships. The node at the start of the arrow is the parent, $pa(i)$, and the node i at the head is the child. For example, in the directed graph, figure 2.3b, node A is the parent of node C and equivalently, node C is the child of node A. The absence of edges in these graphs is significant as it conveys conditional independence between the random variables. It should be noted that these graphs cannot contain any directed cycles. It is possible to determine a factorisation of the joint probability distribution directly from the graph [44]. For example, the factorisation for the directed graph in figure 2.3b is given in equation 2.3.

$$p(A, B, C, D, E) = p(A)p(B|A)p(C|A)p(D|A, B, C)p(E|C) \tag{2.3}$$

In general the factorisation for a directed graph is given by equation 2.4 [44].

$$p(x_{1:N}) = \prod_{i=1}^{N} p(x_i|x_{pa(i)}) \tag{2.4}$$

where $x_{1:N}$ are the N nodes and $x_{pa(i)}$ are the parents of node i

Plates are another useful tool in graphical models, as they make the notation more compact. They are boxes drawn around a group of one or more nodes, representing that the nodes within the box are replicated a given number of times, which is stated in the bottom right hand corner of the box. For example, observations y_n from 1 to N conditioned on A and B can be represented explicitly as shown on the right of

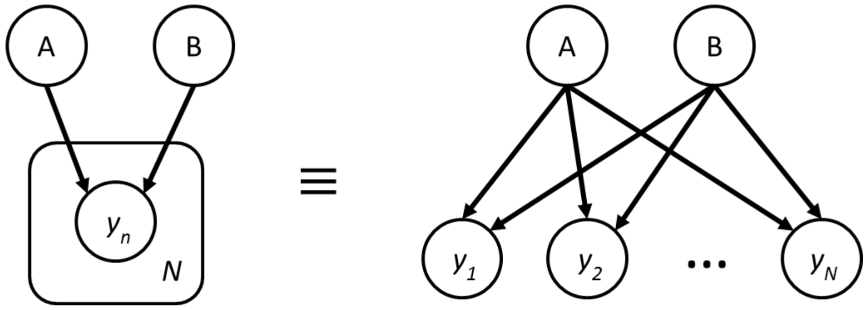


Figure 2.4: Example of plate notation in graphical models

figure 2.4 or equivalently, using the more compact plate notation as shown on the left [44, 45].

The models in figure 2.4 are equivalent to equation 2.5 [44].

$$p(y_{1:N}, A, B) = p(A)p(B) \prod_{n=1}^{N} p(y_n|A, B) \tag{2.5}$$

2.1.3 Topic model concept

Topic models were originally developed to aid the understanding of large corpora of text. They allow hidden topical patterns within a corpus to be uncovered using unsupervised learning. This enables documents to be annotated automatically according to the topics that are referred to in their content. This is particularly useful when using optical character recognition to digitise archived material that has not been given keyword labels. These annotations can be used to organise, summarise and search the whole corpus in new and meaningful ways [36, 41, 46, 47].

In order to discover the hidden topical patterns in a corpus it is required to model the original creation of the documents that make up the corpus. The model does not need to accurately reflect the process undertaken by a human when writing a document but just describe a mathematical procedure that would give the relevant outcome, given a set of assumptions. In particular, it is important to note that the resultant document created using this process is just a bag of words, with no ordering, and hence would be unreadable to a human. The probabilistic generative model of creating a document is shown in figure 2.5. The process involves repeating the following steps for every document in the corpus [41]:

1. Choose a distribution over topics from a Dirichlet distribution (topic proportions for document, represented by the histogram in figure 2.5)

2. For each word:

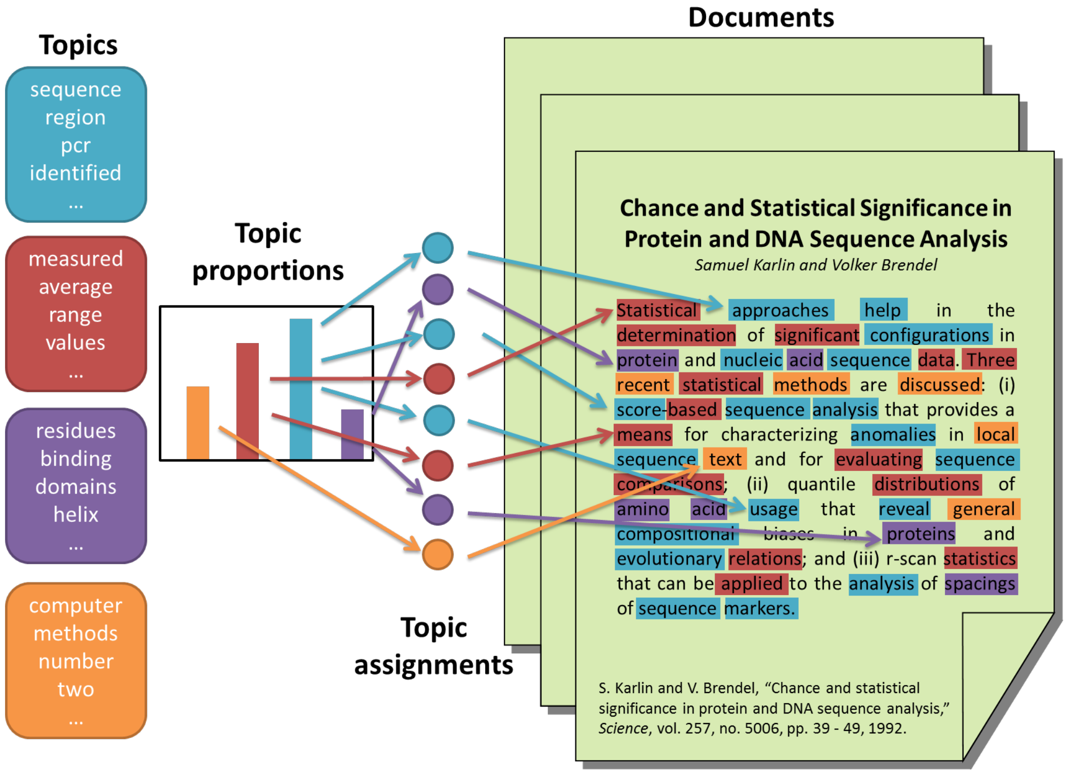


Figure 2.5: Model of probabilistic generative process for creating a document Based on results published in [47]

- Draw a topic from this distribution (topic assignment, represented by a coloured coin in figure 2.5)
- Draw a word from the topic (a distribution over the vocabulary, drawn from a Dirichlet distribution)

The modelling assumptions are [41]:

- The vocabulary of words chosen is fixed for each model
- Topics are distributions over the vocabulary
- The are a fixed number of topics, K, chosen for each model
- The probability of each word varies for each topic
- Each document is a random mixture of corpus-wide topics
- Each word in a document is drawn from one topic

2.2 Latent Dirichlet Allocation (LDA)

This section formalises the mathematical model behind the concept of topic models, known as Latent Dirichlet Allocation (LDA). Different inference methods to approximate the posterior distribution are highlighted. The selected method of variational inference and estimation of the model parameters are then presented in more detail.

2.2.1 Probabilistic model

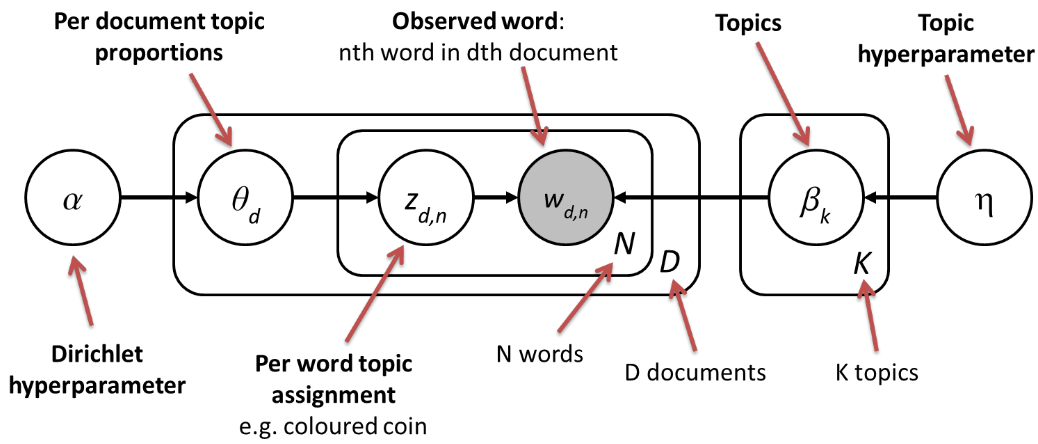


Figure 2.6: Graphical model for Latent Dirichlet Allocation (LDA)

In reality, the process of a document being created is not seen, it is only possible to observe the resultant documents. It is desired to infer the underlying topic structure, that is to work out the distributions that could have been used to create the documents. In order to achieve this it is helpful to model the whole concept as a probabilistic graphical model, known as LDA [46], shown in figure 2.6. The generative process of creating a corpus of documents defines a joint probability distribution over the observed words and the latent variables representing the topics, topic proportions and topic assignments [48]. The joint distribution of all the hidden and observed variables can be determined directly from the graph in figure 2.6 and is given by equation 2.6 [47, 48]. The dependencies between the variables represented by the graphical model and mathematical equation are what defines LDA [48] and match the statistical assumptions behind the generative process described.

$$p(\vec{\beta}_{1:K}, \vec{\theta}_{1:D}, z_{1:D,1:N}, w_{1:D,1:N} | \vec{\alpha}, \eta) = \left(\prod_{k=1}^{K} p(\vec{\beta}_k | \eta)\right) \left(\prod_{d=1}^{D} p(\vec{\theta}_d | \vec{\alpha}) \left(\prod_{n=1}^{N} p(z_{d,n} | \vec{\theta}_d) p(w_{d,n} | z_{d,n}, \vec{\beta}_{1:K})\right)\right) \tag{2.6}$$

where
D is the number of documents in the corpus
N is the number of words in the dth document
K is the number of topics
$\vec{\theta}_d$ are the topic proportions for the dth document
$z_{d,n}$ is the topic assignment for the nth word in the dth document
$w_{d,n}$ is the nth observed word in the dth document and is an element of the fixed vocabulary, which is of size V

$\vec{\beta}_k$ is the kth topic and is a distribution over the fixed vocabulary
$\vec{\alpha}$ are the Dirichlet hyperparameters and η is the scalar topic hyperpararmeter

To establish the hidden topical structure it is necessary to compute the posterior distribution, which is the conditional distribution of the topic structure given the observed documents, as given by equation 2.7 [47, 48].

$$p(\vec{\beta}_{1:K}, \vec{\theta}_{1:D}, z_{1:D,1:N} | w_{1:D,1:N}, \vec{\alpha}, \eta) = \frac{p(\vec{\beta}_{1:K}, \vec{\theta}_{1:D}, z_{1:D,1:N}, w_{1:D,1:N} | \vec{\alpha}, \eta)}{p(w_{1:D,1:N} | \vec{\alpha}, \eta)} \tag{2.7}$$

where $p(w_{1:D,1:N}|\vec{\alpha}, \eta) = \int_{\vec{\beta}_{1:K}} \int_{\vec{\theta}_{1:D}} \sum_{z_{1:D,1:N}} p(\vec{\beta}_{1:K}, \vec{\theta}_{1:D}, z_{1:D,1:N}, w_{1:D,1:N}|\vec{\alpha}, \eta)$ [47, 49]

The numerator is the joint distribution of all the variables, equation 2.6, and can easily be evaluated for any setting of the hidden variables. However, the denominator, which is the marginal probability of the observed words in the corpus under any topic model, is intractable to compute. This is due to the fact that the number of possible hidden topic structures, which have to be summed over to compute this value, is exponentially large [46, 48]. Therefore, it is required to use an approximate posterior inference algorithm to estimate a solution [41, 46].

Various approaches can be taken to perform approximate inference of the posterior, including sampling-based algorithms and deterministic methods. The aim is to estimate an approximation as close to the true posterior, equation 2.7, as possible [48]. Inference methods developed include collapsed Gibbs sampling, mean field variational Bayesian inference, collapsed Bayesian inference and maximum a posteriori estimation. Asuncion et al. compare the similarities and performance of different methods [50]. They demonstrate that the choice of hyperparameters has as large an impact on the results as the choice of algorithm. Although variational Bayesian inference is often the worst performing algorithm, it is not by a significant amount for well chosen hyperparameters. Therefore, for the purposes of this work, David Blei's implementation of variational inference for LDA [51] is used as this allows direct comparison with previous work completed by Huynh et al. [32], as discussed in section 4.1.1.

2.2.2 Variational inference and parameter estimation

The goal of performing inference is to estimate the posterior distribution given the observed words. In other words, to determine the topics, $\vec{\beta}_{1:K}$; per document topic distributions, $\vec{\theta}_{1:D}$ and per word topic assignments, $z_{1:D,1:N}$. The hyperparameters of the LDA model, $\vec{\alpha}$ and η can either be selected in advance or estimated, from an initial seed value, for a given dataset as part of the inference algorithm.

More generally, given a set of observations $x_{1:N}$ and latent variables $z_{1:M}$, it is often difficult to efficiently compute the marginal likelihood of the observations, as is the case for LDA. Variational methods are a deterministic alternative to Markov Chain Monte Carlo (MCMC) methods that use established optimisation techniques instead of sampling. For large problems variational methods are faster and more scalable than MCMC, however they are also biased [41]. The idea is to determine the evidence lower bound (ELBO) of the log probability of the observations, $\log p(x_{1:N})$, given by equation 2.8. Jensen's inequality states that $f(\mathbb{E}[x]) \geq \mathbb{E}[f(x)]$ for a concave function $f(x)$, where $\mathbb{E}$ denotes the expectation and this is used to derive the ELBO.

$$\begin{aligned}
\log p(x_{1:N}) &= \log \int p(z_{1:M}, x_{1:N}) dz_{1:M} \\
&= \log \int p(z_{1:M}, x_{1:N}) \frac{q_\nu(z_{1:M})}{q_\nu(z_{1:M})} dz_{1:M} \\
&= \log \left(\mathbb{E}_{q_\nu} \left[\frac{p(z_{1:M}, x_{1:N})}{q_\nu(z_{1:M})} \right] \right) \\
&\geq \mathbb{E}_{q_\nu} \left[\log \frac{p(z_{1:M}, x_{1:N})}{q_\nu(z_{1:M})} \right] \\
&= \mathbb{E}_{q_\nu}[\log p(z_{1:M}, x_{1:N})] - \mathbb{E}_{q_\nu}[\log q_\nu(z_{1:M})] \qquad (2.8)
\end{aligned}$$

where $\mathbb{E}_{q_\nu}$ is the expectation with respect to q_ν and $p(z_{1:M}, x_{1:N})$ is the joint probability distribution over the observed and latent variables.

It can be seen that a distribution $q_\nu(z_{1:M})$ over the latent variables with free variational parameters ν is introduced by multiplying and dividing by it. These parameters are then optimised to make the bound as tight as possible, in other words, to find a member of the family $q_\nu(z_{1:M})$ that is closest in Kullback-Leibler (KL) divergence to the posterior distribution $p(z_{1:M}|x_{1:N})$ [41, 46, 49]. The factorisation of $q_\nu(z_{1:M})$ determines the complexity of the optimisation.

In mean field variational inference the variational distribution is chosen to be fully factored, as shown in equation 2.9. This means that each latent variable is independent and governed by its own variational parameter ν_i. This is beneficial as it is often the dependence exhibited in the true posterior that makes exact inference intractable [41, 46]. In terms of the graphical model this is equivalent to removing problematic edges between nodes. For example, in LDA the denominator of the true posterior, equation 2.7, is intractable due to the problematic coupling between $\vec{\theta}_d$ and $\vec{\beta}_k$, which arises from the edges between $\vec{\theta}_d$, $z_{d,n}$ and $w_{d,n}$ in the graphical model. The model can be simplified by removing these edges and the $w_{d,n}$ nodes and using free variational parameters for $\vec{\theta}_d$ and $z_{d,n}$ [46].

$$q_\nu(z_{1:M}) = \prod_{i=1}^{M} q_{\nu_i}(z_i) \tag{2.9}$$

Suppose the distribution of each latent variable, z_i, conditional on the observed variables and all other latent variables excluding z_i denoted as z_{-i}, expressed as $p(z_i|z_{-i}, x_{1:N})$ is in the exponential family. Assume also that each factor $q_{\nu_i}(z_i)$ is in the same exponential family. The optimal variational parameters can then be found using the coordinate ascent algorithm, where the update is given by equation 2.10 [52].

$$\nu_i = \mathbb{E}_{q_{-i}}[\eta(z_{-i}, x_{1:N})] \tag{2.10}$$

where η is the natural parameter for a distribution in the exponential family.

Mean field variational inference can be used to estimate the intractable posterior distribution for LDA. Here we consider $\vec{\beta}_{1:K}$ to be a parameter of the observed words, that is a fixed quantity to be estimated, rather than drawing the topic distributions from a Dirichlet with hyperparameter η. The details of this approach are given because the work presented in this thesis is based on Blei's implementation of variational inference for LDA [51] which follows this assumption, as outlined in his seminal paper [46]. Taking this into account, for each document, the variational distribution over the latent variables such that all variables are independent is given by equation 2.11.

$$q(\vec{\theta}_d, z_{d,1:N}|\vec{\gamma}_d, \phi_{d,1:N}) = q(\vec{\theta}_d|\vec{\gamma}_d) \prod_{n=1}^{N} q(z_{d,n}|\phi_{d,n}) \tag{2.11}$$

where $\vec{\gamma}_d$ is the variational Dirichlet parameter for $\vec{\theta}_d$, the topic proportions for the document and $\phi_{d,1:N}$ are the variational multinomial parameters for $z_{d,1:N}$, the per word topic assignments for the document.

The optimal values for the variational parameters are found using the coordinate ascent algorithm. For document d, for each topic k the update equation for $\gamma_{d,k}$ is given by 2.12 and for each word, n the update equation for $\phi_{d,n,k}$ is given by 2.13.

$$\gamma_{d,k} = \alpha_k + \sum_{n=1}^{N} \phi_{d,n,k} \tag{2.12}$$

$$\phi_{d,n,k} \propto \beta_{k,w_{d,n}} \exp\{\mathbb{E}_q[\log(\theta_{d,k})|\gamma_{d,k}]\} \tag{2.13}$$

where α_k is the Dirichlet hyperparameter given the current topic is k; $\beta_{k,w_{d,n}}$ is the probability of the current word in the document $w_{d,n}$ given it is assigned to the current topic k and $\mathbb{E}_q[\log(\theta_{d,k})|\gamma_{d,k}] = \Psi(\gamma_{d,k}) - \Psi(\sum_{j=1}^{K} \gamma_{d,j})$ noting that Ψ is the digamma function i.e. the first derivative of the $\log \Gamma$ function. A full derivation can be found in Blei et al. [46].

The results of performing variational inference at the document level provides a tractable lower bound on the log probability of the observations, as desired. This bound can then be maximised, for fixed values of the variational parameters, with respect to the corpus level parameters, $\vec{\alpha}$ and $\vec{\beta}_{1:K}$. The following alternating variational Expectation Maximisation (EM) algorithm can be used to achieve this. Details of the derivation of this algorithm can be found in Blei et al. [46].

1. (E-step) Find the optimal values for the variational parameters $\vec{\gamma}_d$ and $\phi_{d,1:N}$, for each document and hence an approximation of the posterior distribution $p(\vec{\theta}_{1:D}, z_{1:D,1:N} | w_{1:D,1:N}, \alpha, \vec{\beta}_{1:K})$.
2. (M-step) Find the maximum likelihood estimates for α and $\vec{\beta}_{1:K}$ with expected sufficient statistics for each document under the approximate posterior computed in the E-step.

Repeat both steps until the lower bound converges.

2.3 Extensions to Latent Dirichlet Allocation

Topic models were originally based on Latent Dirichlet Allocation, as discussed in 2.2. However, this model has some short comings and is not suitable for solving all problems. In particular, LDA assumes that data are exchangeable, i.e. the order of documents and order of words does not matter. Furthermore, the use of finite hierarchical mixtures means that the number of topics needs to be known in advance and is then fixed for that model. Several extensions to the original graphical model of LDA have been developed and applied to problems both within and beyond the natural language processing community [36]. Extensions relevant to this thesis are presented in this section.

2.3.1 Dynamic topic models

Dynamic topic models (DTMs) extend LDA by including a model of the evolution of the topics over time, whereas documents are assumed to be exchangeable within a corpus for LDA. This is achieved by dividing the data into T time slices, such as a year, and then modelling the documents in each time slice using LDA with K topics. Each time slice is a separate LDA model and information is only shared between each instance by chaining together the topics and thus sequentially tying the collection of

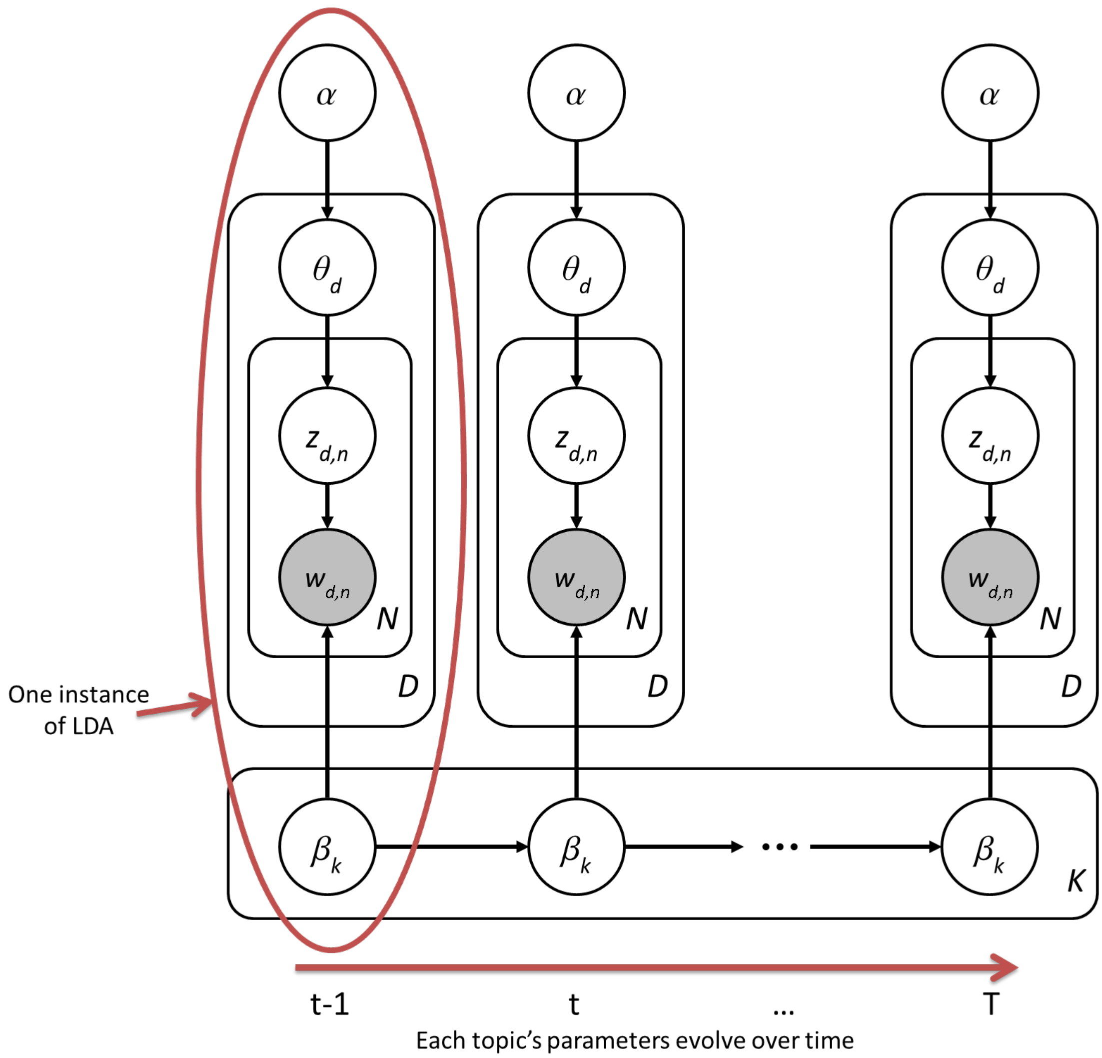


Figure 2.7: Graphical model for Dynamic Topic Model

topic models. The topics can smoothly evolve from time slice $t-1$ to t, as shown in figure 2.7. As the Dirichlet distribution is not suitable for sequential modelling, the logistic normal distribution is used to capture uncertainty about the time-series topics by embedding it in a state-space model. Each topic is represented by a multivariate Gaussian random variable, which are modelled to evolve with Gaussian noise. The words drawn from these topics are then mapped to the simplex [36, 47, 53].

Variational inference can also used to estimate the posterior for DTMs. In fact, the nonconjugacy of the Gaussian and multinomial distributions makes this an easier method to implement than alternative sampling based techniques [53]. The approximate variational posterior for the DTM is given by equation 2.14. Note that at the document level this is the same as equation 2.11 for LDA.

$$\prod_{k=1}^{K} q(\vec{\beta}_{k,1:T}|\vec{\hat{\beta}}_{k,1:T}) \prod_{t=1}^{T} \left(\prod_{d=1}^{D_t} q(\vec{\theta}_{t,d}|\vec{\gamma}_{t,d}) \prod_{n=1}^{N_{t,d}} q(z_{t,d,n}|\phi_{t,d,n}) \right) \tag{2.14}$$

where
K is the number of topics
T is the number of time slices
D_t is the number of documents in time slice t
$N_{t,d}$ is the number of words in document d of time slice t
$\vec{\beta}_{k,1:T}$ are distributions of topic k for each time slice
$\hat{\vec{\beta}}_{k,1:T}$ are the Gaussian variational observations for $\vec{\beta}_{k,1:T}$ that retain the imposed sequential structure
$\vec{\theta}_{t,d}$ is the proportion of topics in document d and time slice t
$\vec{\gamma}_{t,d}$ is the free variational Dirichlet parameter for $\vec{\theta}_{t,d}$
$z_{t,d,n}$ is the topic assignment for word n in document d at time slice t
$\phi_{t,d,n}$ is the free variational multinomial parameter for $z_{t,d,n}$

Hence, the update equations for the document level variational parameters are the same as for LDA, equations 2.12 and 2.13. The conjugate gradient method is used to optimise the variational observations for the topics. An approximation based on a Kalman filter is used in the implementation utilised in this thesis (chapter 5). Further details of this approach can be found in Blei and Lafferty [53].

2.3.2 Other extensions

The two models used in the work of this thesis are LDA and DTMs that have been described. Other extensions to LDA are mentioned here for completeness as they are considered for future work.

LDA requires the number of topics to be defined before estimating the model, however, the choice of the number of topics is not well defined. Often it is chosen by examining the fit to held-out documents or using the marginal probability of the whole collection but this is time consuming and does not always yield the optimal result. Bayesian non-parametric methods can be used to solve this problem. In particular, Hierarchical Dirichlet Processes (HDP) allow the number of topics to be infinite a priori. In addition, previously unseen documents can evoke a previously unseen topic i.e. an atom that has not yet appeared [36].

The graphical model for Hierarchical Dirichlet Processes is shown in figure 2.8. It can be seen that the topic proportions $\vec{\theta}_d$ in LDA, are replaced by a distribution over topics G_d drawn from a Dirichlet Process, $DP(\alpha, G_0)$, where α is the precision parameter and G_0 is the base distribution over topics. G_0 is also drawn from a Dirichlet Process, $DP(\gamma, H)$, where H is a symmetric Dirichlet on the word simplex and γ is the precision parameter. Hence, the atoms of the per-document distributions

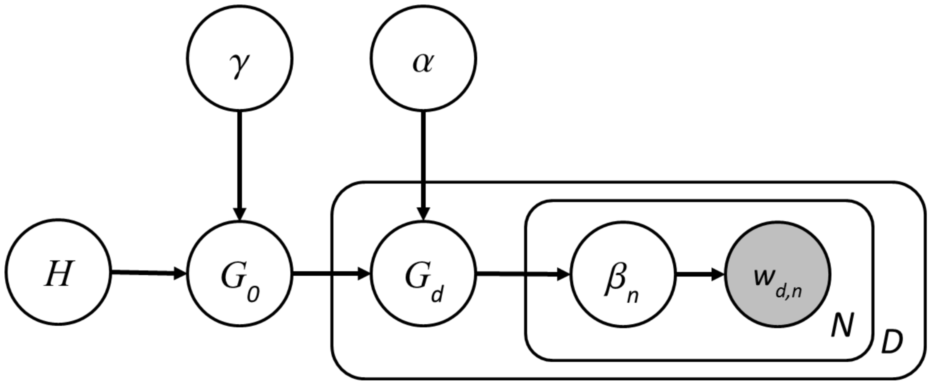


Figure 2.8: Graphical model for Hierarchical Dirichlet Processes

over topics are shared across documents. The clustering property guarantees that the words in a document share a subset of topics [36].

Dynamic Hierarchical Dirichlet Processes (dHDP) extend HDPs by modelling how topic-proportions change over time. This is distinct from a DTM, which models how topics change over time. This is achieved by positing a distribution over topics G_t corresponding to all documents at time t. The model imposes a generally smooth evolution of G_t but allows sharp changes in time if required by the data. As with HDPs, the number of topics is inferred from the data. This model also assumes that data is not temporally exchangeable [36].

The author-topic model [54] extends LDA to model the association between authors and topics using a multinomial distribution. This enables information about which topics authors write about to be obtained, as well as the topics that make up a document. The idea behind the correlated topic model is to assume that there is a relationship between the different topics discovered. This relationship is explicitly modelled using the logistic normal distribution so that it can be quantified as an output of the correlated topic model [47].

2.4 Analysis of topic model performance

Clearly, it is important to analyse the results of topic models to determine how well they are performing [55, 56]. However, this analysis is not straightforward as it is not easy to define what the 'best' performance is. In particular, a model may find patterns that a human would not have thought of but are actually relevant and useful for answering a particular question. A variety of different methods have been developed to try and assess the performance of topic models [55–58]. These can be split into two categories: quantitative and qualitative analysis.

There is no universal method of accurately and efficiently measuring the generalisation of a topic model that is independent of any specific application. However, one method that is often used in the topic modelling literature is an estimate of the probability of held-out documents for a trained model. A measure known as perplexity is the inverse of the geometric mean per-word likelihood on a held-out set of documents using a model learned on a set of training documents. For a set of M test documents the perplexity can be calculated using equation 2.15 [46].

$$\text{perplexity}(D_{1:M}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_{d,1:N})}{\sum_{d=1}^{M} N}\right\} \tag{2.15}$$

where $D_{1:M}$ is the test set of M documents, $w_{d,1:N}$ are the N words in the d^{th} document. A lower perplexity suggests a better generalisation performance of the model because the log function gives high values for low probability words. Moreover, this measure can be used to select the parameters for the model, such as an optimal number of topics for a given dataset [59].

Although quantitative analysis of the performance of a topic model can be useful, it does not necessarily mean that the results are semantically beneficial. In fact, it can often be the case that model A gives more semantically useful results than model B even though model B performs better when measuring the probability of held-out documents [57]. Qualitative analysis investigates the semantic quality of topics found, that is, an expert's judgement of how 'useful' a topic is to the user. Various methods have been proposed to try and quantify the variable semantic quality of topics. These include human scoring of topics [56], word and topic intrusion methods [57] and graph mining techniques [58].

These methods of qualitative evaluation have been developed for the original application of text processing and are not directly applicable to the problems considered in this thesis. They have been used as inspiration for new evaluation methods for the specific applications of activity monitoring and nutrition. In particular, visualisations of topic model results, expert opinions and existing methods from the application areas have been utilised. These are presented in more detail in the relevant chapters.

Chapter 3

Data Fusion for Healthcare Monitoring in Residential Environments

This chapter reviews the literature related to processing and understanding the data that can be collected from a healthcare monitoring system in a residential environment. In particular, a brief summary of the field of data fusion is given for context. Previous work using different types of sensors suitable for this application is considered and gaps in the literature are identified. It should be noted that healthcare monitoring systems focusing on the use of physiological sensors are beyond the scope of this review.

To date, there has been a greater focus on monitoring at the activity level and fewer studies have considered investigating the higher level of behaviour patterns and daily routines that can be detected from activity data [60]. The work that has been done in this area is reviewed in more detail, with a strong focus on the use of topic models to achieve this goal. Moreover, the use of topic models to detect daily routines from other types of data is also considered. Finally, the application of topic models to nutrition data, which could be generated by a residential healthcare monitoring system is addressed. Diet and nutrition are highly correlated with health, especially chronic diseases. Understanding eating behaviour patterns in greater detail aids research into the impact of these relationships.

3.1 Residential healthcare monitoring systems

Monitoring systems in the home can provide individuals with a greater insight into their own health and well-being. They also provide clinicians with previously unavailable data that can aid early diagnosis and treatment of diseases. However, a monitoring system generates a large wealth of data from a variety of sources, which has no clear meaning by itself. In order for the data to benefit the users it must be combined to obtain information that can directly influence decisions and actions [61]. The idea of fusing data from different sensors has been a challenge for a variety of research communities over the last few decades. This section summarises some of the key work in this area and how it relates to this application.

Various types of sensors can be included in a monitoring system, including wearable sensors, environmental sensors and video cameras [62]. This section reviews studies that have investigated activity monitoring in a residential environment, primarily using environmental sensors. The use of wearable sensors for recognition of activities of daily living that occur in a residential environment is also addressed. A 'wearable' or 'body worn' sensor is considered to be a sensor that can be externally attached to a person, such as a wristband, clipped onto a belt or worn as a piece of jewellery. The sensor(s) in the device can vary but for activity recognition accelerometers are the most commonly used [62, 63]. Video based activity recognition is not considered in this review as this was the focus of a separate work package within the SPHERE project.

3.1.1 Data fusion definition and models

The phrase 'data fusion' is used widely in the literature, yet there is still not a consensus on the exact definition of this term and associated expressions [64–66]. This problem was recognised by the remote sensing community and hence a working group was set up to address the need for terms of reference in data fusion, presenting their findings in [65]. It was felt that any definition for data fusion should not be restrictive to a specific domain or method, therefore the following definition was proposed.

> "Data fusion is a formal framework in which are expressed means and tools for the alliance of data originating from different sources. It aims at obtaining information of greater quality; the exact definition of 'greater quality' will depend upon the application"
>
> Buchroithner and Wald, January 1998 [65]

Although various definitions exist, it is clear that the key concepts that underpin the field of data fusion are: combining data; using multiple sources and improving quality [67]. The advantages of using multiple sensors are myriad, as discussed in [68, 69]. They include:

- Inherent redundancy for improved reliability and robustness
- Extended spatial and temporal coverage
- Increased confidence in measurements
- Shorter measurement and response times
- Improved resolution
- Increased dimensionality
- Higher signal-to-noise ratio
- Increased hypothesis discrimination

There are a variety of frameworks and models available in the literature that have been developed to help implement a functional data fusion system [66]. Frameworks provide a structure for designing a complex system composed of multiple data sources which need to be acquired, processed and communicated. If they are not directly applicable they can be modified to meet the specific requirements of a particular application [65, 68, 69].

One of the first and most popular data fusion models was developed by the Joint Directors of Laboratories (JDL) Data Fusion Working Group [70]. Developed in the context of military applications, it has some shortcomings for other application areas but the main principles are fundamental [66, 69, 71]. Moreover, the model has since been extended and used as a basis for other models [68, 70, 72]. The JDL data fusion process model includes three main levels of processing that occur to data which comes in from the sources and is output for human computer interaction. There is a fourth meta-process for refinement of the overall process and preprocessing of raw data is sometimes considered an additional level [70, 73].

Luo and Kay proposed the Multi-sensor Integration Model [74] where integration is the use of multiple sensor information for a task, whilst fusion is considered as the actual combination of data. Both fusion and integration occur simultaneously. The model includes hierarchical, sequential fusion centres and measurements from sensors can be fused at one or more levels [69, 75]. Dasarathy [76] defined the functions of data fusion in terms of the relationship between the type of input and output data, which can be data, features or objects. This categorisation is useful as it can be easily mapped to specific techniques and algorithms [73]. Thomopoulos proposed the use of three modules at different levels: signal, evidence and dynamics level. The levels of fusion can be sequential or interchangeable, depending on the application. Moreover, Thomopoulos highlighted the importance of monotonicity of fused information and robustness against a-priori uncertainty in any data fusion system [69].

More recently several models and frameworks have been suggested for healthcare applications. Lee et al. propose a 'Pervasive Healthcare Architecture' [77] which includes a data fusion module to combine data from different types of sensors used in a pervasive healthcare monitoring system. Gong et al. [78] incorporate multiple preferences in their model to provide flexibility in what is considered important in different applications. The model is suitable for many wireless sensor networks but is designed with a specific focus on healthcare monitoring. A customisable framework for collaborative body sensor network applications is described in [79]. The framework supports a three-layer model for multi-sensor data fusion, which is extended to support collaborative computing between networks.

Rodriguez et al. [80] present a system based on PANGEA, a multi-agent architecture which provides a high level framework for intelligent information fusion. They demonstrate the use of this healthcare monitoring system in a residential care home for 8 months. King et al [34] describe a generic centralised hierarchical data fusion model for wearable health monitoring, incorporating ideas from the JDL and Multi-sensor Integration models. This model uses three levels of abstraction, signal, feature and decision level fusion as well as pre-processing and feature extraction steps as required. Data from each sensor passes through one or more of the fusion centres as appropriate. Cao et al. [81] developed an integrated framework for multi-modal activity recognition that includes correcting for class imbalance, caused by varying frequencies and durations of different types of activities. They propose that fusion happens at the post classification level, i.e. the labels predicted by different classifiers are fused to reduce errors.

3.1.2 Data fusion algorithms

There are many different data fusion algorithms and techniques [66]. These are often categorised into three levels, although the names given to each level varies. Here, the simple labels of low, middle and high level fusion are used to minimise confusion. Luo et al. [75] provide a useful reference table that summarises the key advantages and disadvantages of some of the methods for each level. An alternative approach to looking at available data fusion algorithms and techniques is taken by Khaleghi et al. [66], where methods are considered in terms of data-related challenges that can be addressed. In particular, different frameworks for dealing with uncertainty are highlighted. Moreover, Castanedo [67] outlines the different approaches that can be taken to classifying data fusion techniques. A summary of the type of algorithms and specific examples based on the three fusion levels convention are given here for context. The choice of algorithm relies on finding an agreement between the application requirements and the relevant advantages and disadvantages of available techniques at the appropriate level.

At the low level of data fusion there is a need to synchronise and adapt the data due to varying sampling properties across multiple sensors. Techniques for low level fusion can be broadly separated into two categories: statistical estimation and covariance based methods. Statistical estimation methods can be non-recursive, such as weighted averages or least squares, which both merge redundant data. Alternatively, they can be recursive, for example a Kalman Filter, which is a predict-update type of estimator. Furthermore there are also Extended Kalman Filters, which linearise about a point and Unscented Kalman Filters for non-linear filtering. Covariance based methods include cross covariance, covariance intersection which is used if sensor data are not independent and covariance union which solves the problem of information corruption [75].

The aim of the middle data fusion level is to group the data into classified sets within a multidimensional feature space. Again, techniques can be broadly separated into two categories: parametric and non-parametric. Parametric methods include parametric templates, which are used for image processing, k-means clustering and Naïve Bayes. The non-parametric methods include learning vector quantization, kohonen feature maps, artificial neural networks and support vector machines. Prior assumptions about the distribution of input data are not required for non-parametric classification algorithms, which can be advantageous for some applications [75].

High level data fusion methods use a symbolic representation of information. Symbols, with an associated level of uncertainty, are combined to give a composite decision. Therefore, algorithms at this level must be tolerant of imprecision and uncertainty. There are a variety of techniques that can be used for inference. Bayesian inference is a probabilistic framework for recursive state estimation. Particle filters are a modern sequential monte carlo bayesian method based on point mass representation of posterior probability density. Topic models are a type of hierarchical mixture model, used to find hidden patterns [36]. Dempster-Shafer theory is based on the concept of degrees of belief whilst fuzzy logic is a multivalue logic that allows uncertainty in multi-sensor fusion to be directly represented. Artificial intelligence methods, such as neural networks, can also be used for high level fusion [75]. Each method has associated modelling assumptions which must fit with the problem domain they are used for.

Grouping available algorithms in different ways is a useful reference point when selecting algorithms, however expert knowledge about the specific application is still required to choose the most appropriate technique. Various review papers have presented different view points for considering appropriate data fusion techniques within the broad fields of healthcare monitoring and activity recognition. Lai et al. [82] present a survey of body sensor networks, which are often used for healthcare monitoring. They include a section that reviews the data fusion methods that can be used for both physiological sensors and inertial sensors. Major challenges in the field are highlighted, including the impact of architecture choice and data management requirements on the selection of fusion methods. King et al. [34] concentrate on data fusion techniques for wearable health monitoring systems. A discussion of the considerations for selecting a suitable algorithm for a particular application is included.

Pires et al. [83] focus their comprehensive review on fusion techniques for identifying activities of daily living within the constraints of using mobile devices. They include a summary of advantages and disadvantages of different categories of methods, but conclude that the choice of the best method is dependent on many factors specific to a particular application. The challenges related to data fusion for fall detection systems, which can be incorporated in a more general healthcare monitoring system, are considered by Koshmak et al. [84]. In particular, they emphasise the benefit of developing a system with more than one sensor type e.g. ambient and wearable. In addition, they propose that monitoring systems should be modular so that different components can be added and removed as required.

A detailed review of sensor-based activity recognition is given by Chen et al. [63] with application to many areas, including healthcare. They review activity recognition models and algorithms using a hierarchical structure, firstly differentiating between data-driven and knowledge-driven approaches and then sub-dividing these categories further. Within the data-driven category, generative and discriminative models are considered and mining, logic and ontology based approaches are included in the knowledge-driven category. The advantages and disadvantages of these different approaches are discussed and summarised. Moreover, the impact of the context on the choice of approach is highlighted. Future research trends are considered, including the use of domain knowledge and context ontologies to facilitate high level semantic data fusion. Other surveys discuss algorithms used for activity recognition at different levels but do not specifically focus on data fusion. Thus, these papers are presented in section 3.1.4.

3.1.3 Activity monitoring in a residential environment

The literature in the field of activity monitoring in a residential environment is vast and it is not feasible to include everything here. Ni et al. [62] provide an extremely comprehensive survey of various aspects of smart home technology, with a particular focus on older people's independent living. They dedicate section 2 to reviewing recent surveys, projects and applications of smart homes. Alam et al. [85] discuss the history and future of smart homes, including the application of healthcare. They consider a wide range of aspects from communication protocols to sensor types and algorithm choice as well as reviewing relevant projects. Detailed reviews of both Ambient-Assisted Living (AAL) tools and smart home sensor technology are provided by Rashidi & Mihailidis [86] and Ding et al. [87] respectively.

This section only considers activity monitoring within a real residential environment for a period of greater than 24 hours. Data collection of activities in a simulated residential environment set up within a laboratory are not included here. This focus was chosen as it is in-line with the work of the SPHERE project, with which this PhD was associated. Applying different algorithms and techniques to data collected 'in the wild' can produce very different results compared with controlled laboratory conditions due to the unpredictable nature and challenges of a complex, natural environment [88]. Furthermore, short snapshots of data do not provide a realistic representation of the variety of activities that a person performs.

Several studies have investigated monitoring of people using sensors installed in their own home [89–93] or a dedicated smart home / AAL environment [26, 27, 94] for healthcare and well-being applications. These studies primarily use environmental sensors within the home to detect users' activities. The sensors used include: various motion detectors, including infra-red sensors [26, 89–91, 93, 94]; reed or magnetic switches on doors and cupboards [27, 89, 91, 94]; toilet flush and water flow sensors [27, 91]; watt-meters or electric current flow sensors [27, 89]; environmental sensors e.g. humidity, light, barometric pressure [27]; gas sensors [27]; stove sensor (to detect heat or flames) [26, 89, 93, 94]; motion sensors mounted on specific objects, including a bed [26, 27, 93, 94] and a gait / fall monitor based on floor vibrations [93]. A few key projects are presented to highlight the benefits and limitations of different types of sensors and system designs.

One significant project is the PlaceLab at Massachusetts Institute of Technology, a highly instrumented living environment, consisting of a living room, dining area, kitchen, small office, bedroom and bathroom, that can be used for observing activities and interactions of everyday home life [95]. The environment includes an addressable speaker system and a large network of sensors that can be added to as required.

Sensors include switches on doors and cupboards, environmental sensors, water and gas flow sensors, cameras and microphones. Additionally, accelerometers can be attached to moveable objects and worn by participants living in the apartment. A dataset of 104 hours of annotated data from over 900 sensors of a couple who lived in the PlaceLab for 10 weeks was collected [27]. The annotations were done by a third party based on the video data. A detailed discussion of the limitations and challenges of performing activity recognition on data from a real-world setting is provided in [27]. In particular, complex behaviours, such as multi-tasking and interactions between participants make it challenging to differentiate between activities and label them effectively. Using a large number of sensors integrated into a dedicated living space allows an extremely rich and detailed dataset to be collected. However, the cost of installing and maintaining complex systems is not feasible on a large scale and it is intrusive for users to have so many sensors, especially cameras, in their living space.

The CASAS (Center for Advanced Studies in Adaptive Systems) project at Washington State University [96] has a dedicated three-bedroom smart apartment on their campus. The apartment is instrumented with a range of sensors, including infra-red motion/light detectors, ambient temperature sensors, door switches and sensors for monitoring water, stove and phone use. This testbed is often used for validating new algorithms but can also accommodate residents living there for longer term studies. A novel activity discovery method was developed that uses an unsupervised approach to recognise and track activities in the smart home. This was applied to three months of daily activity data collected from two residents [96]. Unsupervised approaches offer the strong advantage of not requiring annotated training data which is challenging to collect.

A large number of sensors, such as used in the PlaceLab and main CASAS test bed, is not viable at scale. The CASAS project developed a ‘smart home in a box’ solution for easily instrumenting residential environments with smart home technology. The system is affordable for large scale deployment and can be installed within 2 hours without making any changes to the home. It includes motion detectors, door and temperature sensors. The CASAS design has currently been installed in 32 testbeds and datasets collected from these environments are publicly available [94]. This is an invaluable resource for testing and comparing different methods, with greater confidence in the conclusions. The disadvantage of reduced scale systems is the lack of detail in the data collected and challenges in detecting anomalies, such as visitors.

Van Kasteren et al. [97, 98] have also designed an easy-to-install system consisting of contact switches, pressure mats, mercury contacts, passive infrared sensors and float sensors for the toilet flush. The addition of extra sensors such as the float sensor do not significantly increase the cost and complexity of the system but can add extra important detail and context. For example, only a motion sensor in the bathroom does not give any information about what someone is doing, whereas knowing the toilet has been flushed means it is highly likely that the toilet has been used and this can be a useful measure for health outcomes. This system was installed in three different residential environments of varying size and layout and hence different numbers of sensors.

Datasets were collected for different users, living in each residence over a period of 2 - 4 weeks. These were annotated using either a handwritten diary or a Bluetooth headset and speech recognition software. The datasets and source code for the probabilistic models used to analyse the data have been made publicly available. The variety in the datasets collected enables transfer learning to be applied when little or no labelled data is available for a new house by using a model learnt on a previous house. Ordonez et al. compared a full Bayesian inference approach with learning the maximum a posterior model parameters when using a hidden Markov model (HMM) for activity recognition [99]. They demonstrated that the Bayesian inference gives a significantly better performance, particularly in a transfer learning framework.

The University of Virginia developed a wireless in-home monitoring system that consisted of motion sensors, a stove temperature sensor and a bed sensor system [26]. The system was installed in the assisted living units of 22 participants, 7 of whom were memory care unit residents, for a period of 3 months. Collecting data for a relevant cohort of participants is vital to ensure that the systems and algorithms developed are fit for purpose. A rule-based inference method was used to automatically determine the participants key activities of daily living (ADLs). The results from multiple users were used to prioritise the care needs of the patients [26]. Another study installed the same system, with the addition of a passive floor-vibration gait monitor and fall detector, in the homes of 13 home health care agency clients for four months. One particular case study highlighted that the system's ability to monitor ADLs, such as bathroom visits could have avoided a patient's hospitalisation had the doctors been able to see this data [93].

The Empath platform was also developed at the University of Virginia [100] to monitor people with depression. This platform integrates both the analysis software and the hardware, including: motion sensors and contact reed switches for activity detection; a sleep monitoring system using accelerometers in the bed; WiFi weighing scales; a microphone on a touchscreen device for speech analysis and an iPhone to record mood. The Empath system was installed, in less than an hour, in the apartment of a patient for a period of 14 days. The study demonstrated that various factors linked to depression could be successfully measured in a home environment. It shows how sensors can be selected to provide specific data that is relevant for monitoring a known condition. However, the activity detection could be improved to give more detailed information than room occupancy and estimated energy levels.

3.1.4 Wearable sensors for activity recognition

The focus for this section is to review the literature in the area of detecting activities using wearable sensors. They offer the advantages of being directly associated with a specific person and providing more detailed information about an individual's movements and activities. User compliance is critical though, as no data can be collected if the user forgets to wear the device, which is a particular risk for certain cohorts, such as people with dementia. Human activity recognition has a wide range of applications and includes detection of simple actions or gestures, through to higher level, more complex activities such as 'having dinner' [63,101,102]. Chen et al. [60] present an overview of recent survey and review papers for activity recognition, focusing on inertial sensors in section 3. This highlights the vast nature of the literature, which is summarised in this section, including the fusion of wearable sensors with other sensor types. Table 3.1 provides a reference for the different areas of focus for the wearable sensor review papers discussed here.

Table 3.1: Summary of areas of focus for review papers on wearable sensors for activity recognition

Focus	**Reference**							
	[101]	**[103]**	**[104]**	**[105]**	**[106]**	**[107]**	**[108]**	**[109]**
Feature extraction	✓	✓	✓	✓	✓			✓
Classification algorithms	✓	✓	✓	✓		✓	✓	✓
System analysis	✓	✓	✓					
Healthcare application			✓	✓			✓	
Sensor types	✓	✓	✓					✓
Sensor placement	✓	✓	✓	✓				

Bulling et al. [103] provide a comprehensive tutorial on human activity recognition with a specific focus on the use of on-body inertial sensors. Godfrey et al. [104] give a detailed background on the use of accelerometers to measure human movement, including considerations of sensor placement and algorithms for data analysis. Lara and Labrador [101] survey the state of the art for a wide range of aspects of human activity recognition using wearable sensors: from system design to classification algorithms and evaluations of complete systems. They present a taxonomy to assist the comparison and analysis of different systems that have similar characteristics.

Finally, they highlight important directions for future research efforts, such as addressing the complex nature of composite and concurrent activities, including at a routine level. A survey by Avci et al. [105] analyses research in activity recognition using inertial sensors with applications in healthcare, well-being and sports. They segment the activity recognition process into five steps and present the algorithms and techniques that can be utilised at each stage. Classification of individual activities is the highest level of fusion included in this survey. They include a summary of the relationship between different classification techniques, sensor placement and the corresponding reported accuracy level.

Two review papers in this area have been contributed by Preece et al. [106, 107]. The first focuses on comparing 14 methods, based on wavelet transforms and frequency and time characteristics, for extracting classification features from accelerometer data for recognising activities [106]. They used two datasets to compare the different feature sets used as input to a k-nearest neighbour classifier. It was shown that time and frequency characteristics based on an ankle sensor outperformed wavelet transforms for dynamic activities carried out by healthy participants. However, this is a limited conclusion because the results will vary for different classifiers and demographics. The second paper provides a review of classification techniques for activity identification using wearable sensors [107]. This includes techniques at different levels, from windowing the raw sensor data through to classification algorithms. They highlight studies that use a range of different machine learning techniques for activity recognition and also focus specifically on fall detection. They emphasise the variability in the nature of datasets, making meaningful comparisons between methods challenging.

Jatoba et al. [108] review a variety of techniques for classification of physical activity as a method to provide context when monitoring the ECG (electrocardiogram) and blood pressure of a patient with cardiovascular disease over a long period in a non-clinical environment. Different levels are considered, from data preprocessing through to classification algorithms. They collected a dataset of daily activities and compared the accuracy of different methods. For this specific dataset it was found that the best results were given by the classification and regression tree and adaptive neuro-fuzzy inference system. In addition, Mannini and Sabatini [109] also review a range of activity recognition methods and describe their application to simulated datasets based on the daily activity accelerometer dataset from Bao and Intille [110]. They compare the results and conclude that a continuous hidden Markov model is optimal when spurious data is removed for this dataset. Moreover, they summarise a variety of human activity classification systems in the literature, including details of sensors, features, classifiers, number of activities, number of subjects and accuracy.

Seminal work by Bao and Intille in 2004 [110] used 5 bi-axial accelerometers, with a sampling rate of 76.25 Hz to collect data from 20 participants performing a range of 20 activities. These were selected to reflect common everyday activities at a variety of intensity levels and included, reading, walking, brushing teeth and folding laundry. Both a semi-naturalistic collection protocol and a more controlled collection procedure were used for the experiments. Mean, energy, entropy and correlation features were created from the accelerometer data. These were used with decision tables, instance-based learning (nearest neighbour), C4.5 decision trees and naive Bayes classifiers to perform activity recognition. It was found that the highest recognition accuracy was achieved with the decision tree classifiers, with an overall accuracy rate of 84%, followed by the nearest neighbour algorithm. The results also demonstrated that a high recognition accuracy, comparable with previous works in a controlled laboratory environment, was achieved with data collected using the semi-naturalistic protocol.

Using multiple sensors is not always realistic for real-world long term monitoring as users can find the devices obtrusive. In contrast to Bao and Intille, Wang et al. [111] used a single waist worn tri-axial accelerometer with a hidden Markov model (HMM) based method to classify six different human daily activities. Data was collected, at a sampling rate of 50 Hz, for 13 healthy subjects, aged 26 to 50 years, who were asked to perform the following activities: falling, jumping, running, sitting down, standing and walking. A subset of the data collected from the accelerometer was used to train a HMM for each of the six activities. Different parameters were investigated and tested on the unseen data. It was found that the highest recognition accuracy of 94.8% was achieved with 7 hidden states and 3 mixture components. Furthermore, it was found that the most likely activities to be misclassified are sitting and falling, this is likely to be because from the point of view of accelerometer data sitting is similar to a controlled fall.

Combining data from wearable sensors with other sources of information, such as cameras and motion detectors can improve the accuracy of activity recognition [60,112]. Peetom et al. [112] conducted a systematic document search of key databases for articles related to monitoring technologies for detecting activities of daily living and significant events, with a specific focus on independent living of the elderly in their homes. They identified five types of monitoring technologies, including the use of body-worn sensors. They summarise the 26 articles found in this category in a table, focussing on the aim of the monitoring, study characteristics and main outcomes. Furthermore, they also found 31 articles that utilise a combination of the five technologies highlighted. In particular, 5 papers combined body-worn sensors and passive infra-red (PIR) motion sensors, this was the second most frequent combination after video and PIR. However, the majority of these papers described validation studies in a laboratory environment.

Although it did not match the criteria of the search conducted by Peetoom et al., one key project that combines different types of sensors, not yet discussed in this review, is the Opportunity Project [113]. They used a highly instrumented environment to collect a dataset of complex, naturalistic activities of daily living. The environment was set up in a room and designed to mimic a studio flat. Participants were asked to follow a high-level script of activities but were free to interpret and perform them as they would normally, to capture natural variations between participants. A total of 72 sensors of 10 different modalities were utilised, including wearable sensors, environmental sensors and sensors attached to specific objects of interest, such as a knife. A subset of the collected data was released as an open source dataset that can be used for benchmarking different approaches and techniques [114]. In particular, as part of an activity recognition challenge that was held in 2011 based on this dataset, baseline performance of standard classification techniques were provided [114]. These were k-Nearest Neighbour, Nearest Centroid Classifier, Linear Discriminant Analysis and Quadratic Discriminant Analysis. The tasks in this challenge included multimodal activity recognition of 4 modes of locomotion e.g. walking and 17 gestures e.g. open drawer.

Finally, the SPHERE project, with which this PhD was associated, held a similar challenge in 2016 [115]. However, the data was collected in a real house environment rather than simulated. Participants were asked to follow a set of scripted activities and also record naturalistic activities that were not scripted. The dataset included raw signals from a wrist-worn accelerometer and PIR sensors, as well as bounding box and centre of mass features from RGB-D cameras (no raw video data was used for privacy reasons). The main aim was to predict posture (e.g. sitting), transition (e.g. sit to stand) and ambulation e.g. (walking) labels for every second of the dataset. A total of 20 labels, across all 3 categories were used to annotate the data. The winning competition entry [116] used a location-based hierarchical approach and temporal activity consistency for classification. This performed better than the baseline features, with a micro-f1 score of 0.76 reported. However, this approach may not perform as well on non-scripted activities as there is likely to be more variation in location-activity relationships and the temporal order of activities recorded in a free-living scenario.

The use of wearable sensors for activity recognition has received a lot of attention from a variety of research communities. Advances in low-power, low-cost and miniaturised sensors and improvements in wireless communication technologies have made the application of wearable technology a reality. The quantity and complexity of data produced presents significant challenges in terms of exploiting its full potential.

Real world systems have many requirements that often constitute a direct trade off. For example wearing multiple sensors can provide a much more accurate picture of a person's activities and movement but are an inconvenience to wear and recharge. If the extra level of detail gained makes a significant improvement on the quality of life for an individual with a complex, chronic disease the benefits can outweigh the disadvantages. However, for others simply measuring activity intensity from a single wrist worn device may be sufficient. Similarly, the choice of algorithms depends on the overall data fusion architecture selected, the accuracy required, power consumption and processing time. These choices have to made in the context of the target application. Data fusion from different types of sensors can help to find suitable compromises by combining the advantages of different modalities to provide additional context and detail.

3.2 Detecting behaviour patterns in activity data

Section 3.1 demonstrated that there has been a large amount of research into analysing data that could be generated by a residential healthcare monitoring system with the goal of recognising various activities. The level of granularity at which these activities are detected varies from very short, specific gestures to longer, more complex activities. However, they are all at a lower level of data fusion than behaviour patterns and daily routines. Detecting these higher level patterns facilitates long term monitoring and identification of changes in behaviour or health status. This can aid early diagnosis, enable more effective management of chronic conditions and improve quality of life [3, 11, 19, 24, 26]. In comparison to activity recognition, there has been little research in higher level patterns and changes over time, which many authors have suggested as important avenues of future research [62, 63, 101, 102, 112, 117, 118].

Daily routine discovery is looking for higher level patterns in a person's daily activities that correspond to their routine. For example, a lunch routine could consist of: walking, preparing food, cooking, eating and washing up. Each routine involves a combination of different activities in varying proportions. Some activities will be very specific to a particular routine, such as working at desk, which is most likely to be in the work routine. Other activities, such as using the toilet will occur in many different routines. There are several methods for detecting someone's daily routine from various types of data. Previous work reported in the literature, is highlighted and discussed here.

3.2.1 Using topic models to detect routines

One method for detecting daily routines is to use topic models (Latent Dirichlet Allocation (LDA) and variations), which were introduced in chapter 2. In order to use topic models for an alternative purpose the concept of the generative model needs to be translated to the new application domain. For instance, the daily activities, such as eating, can be equivalent to the words in a document and the routines, such as lunch, equivalent to the topics. For continuous data, the documents are usually specific length time slices, such as 30 minutes, and the entire dataset is equivalent to the corpus. As for activity recognition, the source of sensor data used with topic models for discovering daily routines can vary. This section reviews the work done in this area using (global) location data, wearable sensor data and environment sensor data, including local location data within a building. Studies based on data collected from video cameras is not considered here, as this is not the focus of this thesis.

3.2.1.1 Location data as input

A person's daily routine can be detected from data about their location over time. For example, Ferrari and Mamei [119] used data from Google Latitude, a mobile phone application that collects location information, to look at patterns in people's global movement. Data was collected for 2 people over nearly a one year period and preprocessed to discover and label relevant places for each user, such as home and work. Using these labels and a label to indicate a time-slot within the day, a bag-of-words representation of the data was created and used as the input to LDA. The resulting topics were analysed and shown to be effective at identifying users' routine behaviours in terms of locations.

Farrahi and Gatica-Perez [120, 121] also consider using location data from mobile phones to discover daily routines. They use a much larger dataset, the Reality Mining Dataset from Massachusetts Institute of Technology (MIT), with 97 users over a period of 16 months, although some days are discarded due to having no reception. They use a similar bag-of-words representation based on specific locations, as done by Ferrari and Mamei. In addition to discovering location-driven routines using LDA, they also use Author-Topic Models, described in section 2.3.2, to identify how routines vary between users. The routines found from this type of input data, such as 'going to work late', are not very specific as this simply summarises a transition between locations during a given time range. In comparison, a routine such as 'commuting to work', summarises more detailed information e.g. modes of transportation, about the activities involved, for example 'walking', 'waiting for bus' and 'travelling on bus'.

The more generic routines are used to determine overall behaviour patterns in users and groups of users but only reflect significant changes in geographical location.

A topic model variant was also applied to the MIT Reality Mining dataset more recently by Nabaei et al. [122]. However, they focus on robustly detecting anomalous patterns, with a focus on security and fraud applications. Topic models are learnt for each behaviour and the results are used as a user profile which represents normal behaviour. Gradual changes are distinguished from abrupt ones and the model parameters are updated to reflect the normal variation over time. The model was also applied to a dataset of Brazilian medical hospital records to detect fraudulent insurance claims. The results demonstrated that this method performed well for datasets with a large number of unique attributes.

Furthermore, Steinhoff and Schiele [123] use data from GSM (Global System for Mobile Communications) and Bluetooth as an input to LDA to discover daily routines. It is found that although topics can be discovered from this data, Bluetooth environments are unstable and GSM traces are too static for long periods, hence it is likely that more interesting results would be obtained when combining this with other sensors. Although the ideas from these studies can be applicable, the data sources considered are not suitable for smaller geographical areas, such as in a residential environment, as required for the SPHERE project. Moreover, they do not provide detail of the specific activities included in the user's routine and hence do not directly add useful information for healthcare monitoring applications. They could be used in conjunction with other techniques to provide context, such as how often someone eats out or visits the gym.

3.2.1.2 Wearable sensor data as input

Huynh et al. [32] used topic models to discover daily routine patterns in accelerometer data. They collected 7 days of data from one person wearing two accelerometers. The user annotated their daily routines and the detailed activities that made up these routines. A hierarchical approach was taken to analyse the data, with both supervised and unsupervised methods. Firstly, the raw sensor data was classified into activities. The classifier results were then used as the input to a topic model by creating documents of 30 minutes of activity data. The results showed that the topic models were able to identify topics that correlated well with the ground truth daily routine labels. Further details of the collection and analysis of the dataset (UbiComp08) are given in section 4.1.

Seiter et al. [124] investigated how different characteristics of a daily activity dataset can affect the performance stability of a topic model. They simulated datasets with specific properties based on the original UbiComp08 dataset collected by Huynh et al. A three layered simulation model was defined to sample the daily routines and the activity primitives of which they are composed. The different datasets were used as an input to LDA and the results compared. As the number of topics does not necessarily match the number of routines, a k-Nearest Neighbours classifier was used to map the topics to routines. Details of the key dataset properties found for stable performance of the topic model are discussed in section 4.2.1.

Furthermore, Seiter et al. [125] also considered the performance of three variations of topic models on different datasets, including UbiComp08, benchmarked against a baseline method using k-means clustering. The approaches included LDA, n-gram topic models, which add primitive sequence information and correlated topic models, which explicitly model the relationship between topics, as discussed in section 2.3.2. Seiter et al. define four properties that describe the complexity of a dataset; averaged primitive rate, activity composite specificity, activity primitive sequence similarity and composite-instance ratio. The three datasets analysed all contain activity primitives and composites, however they vary in terms of their properties. After pre and post processing the results showed that overall LDA had the best performance on all datasets, even when noise was introduced. Moreover, it was noted that LDA had a greater advantage over the baseline for datasets with a lower activity composite specificity. This is likely to be due to the underlying assumption of the LDA model that documents exhibit multiple topics. The results also demonstrated that good choices for segment size and the number of topics are close to the weighted mean durations and expected number of activity composites respectively.

More recently, Seiter et al. [126] evaluated the use of LDA for daily routine discovery in hemiparetic rehabilitation patients. Data was collected for 11 patients in a daycare centre using six wearable sensors, although only data from three were required for analysis. Ground truth labels were annotated for six daily routines for each patient for up to 10 days. An unsupervised hierarchical approach was used to detect the routines. The activity vocabulary used as input to LDA is created using either k-means clustering or a rule-based approach based on features extracted from the raw sensor data. The final step to map the discovered topics to the ground truth labels for evaluation uses a supervised k-nn classifier. The results revealed that the rule-based approach performed best, with an average accuracy of 76% across patients. This suggests that the inclusion of expert knowledge is beneficial. However, this may be due to the structured scenario of a rehabilitation centre where specified routines are of interest. In a free-living home environment, the discovery of new routines that have not been identified a priori could provide as much or more insight into the patient's health status than known routines. For example, the model may identify a routine where the user has an increased level of sedentary or active behaviour in comparison with their usual behaviour, that they would not have considered as a routine if asked. By discovering this routine and understanding why the patient was not conscious of it can provide an insight into how to approach changing the patient's behaviour to improve their health and well-being.

Spina et al. [12] conducted a similar study to Seiter et al. [126] but for patients with chronic obstructive pulmonary disease (COPD). However, their dataset was much larger, including 977 COPD patients and 66 healthy controls from studies across 10 countries worldwide. The physical activity of participants was recorded using a commercial wearable sensor which provides estimates of energy expenditure, metabolic equivalent of tasks, step count and sleep status for at least 4 days. Moreover, Spina et al. took a different approach to constructing the vocabulary, using a complex data-driven methodology based on activity intensity categories (IC) and relevant features such as skin temperature, sleep count and accelerometer data. The resulting routines were distributions over the vocabulary and were interpreted by considering the IC to which they were most related. The results showed that the probability of routines inferred over a day varies between different cohorts of patients with varying levels of disease severity. These routines have the potential to be used as biomarkers in diagnostic decision support systems and daily feedback from a monitoring system. The results of this approach are well aligned with the target application but do not provide any information about the activities performed and hence are not necessarily transferable to alternative conditions with different requirements.

Kim et al. [127] used low level activities found from multi-modal sensor data, accelerometer and ECG, as input to LDA. They investigated the performance with the KNOWME real-life dataset, which includes self-labelled high level activities and the corresponding sensor data for 12 subjects in free-living settings. The results reported show that, for a model with 30 topics, all of the labelled high level activities are assigned to only one or two topics. This suggests that this method is not suitable for this dataset. It is likely that this is due to several factors, including the limited number of low level activities (9) and short duration used to create documents (10 mins) in comparison to the reported durations of self-labelled activities. This would mean that the input does not provide enough detail to discriminate between different types of high level activities and documents would be unlikely to exhibit multiple topics as assumed by LDA.

A combination of wearable sensors, including accelerometers and physiological sensors, were used by Peng et al. [128] to find complex activities following a hierarchical approach. They used k-means clustering on features from the acceleration data to create words as input for LDA. The resulting topics were fused with the physiological features at the classifier level. A modified stacking framework method was applied with the J48 decision tree as a base classifier and multinomial logistic regression for the meta classifier. The authors compared their results with a range of other methods, including the original work by Huynh et al. They demonstrated that their method outperformed the others overall in terms of the F-score. However, this is likely to be because the parameters they used were chosen to be optimal for the proposed method. In particular, they applied the approach of Huynh et al. using only 8 low level activity labels that are not well aligned to discriminate between the higher level activities. Moreover, the final stage of this method relies on supervised learning and hence labelled data is required for training, which is challenging to obtain.

Sun et al. [129] developed a two phase non-parametric framework for human routine discovery. The first phase creates artificial words from features using a Dirichlet process Gaussian mixture model, allowing the size of the activity vocabulary to be found automatically. These words are grouped into documents and input to the second phase, a hierarchical Dirichlet process (HDP), enabling the optimal number of latent topics to be discovered by the model. The resulting document topic proportions are clustered using the affinity propagation algorithm to find high level routines, without specifying the number of routines in advance. The framework was tested on two public datasets: the UbiComp08 dataset and a transportation mode dataset from Microsoft Research Asia. The results demonstrated that the non-parametric framework performed comparably to parametric models, such as LDA, with the advantage of not needing to choose parameters. This is useful because there is no obvious way to select the optimal parameters.

A non-parametric approach to routine detection was also considered by Nguyen et al. [130], with a focus on the intensity of physical activity based on accelerometer data. They use a HDP to infer the number of physical activity levels automatically from the accelerometer data. The discovered mixture proportions of activity levels are used as features for a multi-label classifier. The resulting classifications are then compared to the daily routine of the user. In order to test the framework a new dataset was collected for 13 people during working hours over a period of 3 weeks, using the accelerometer in the Sociometric Badge, a device originally developed by the MIT Media Lab. Experience sampling was used to collect the ground truth labels, as it is thought that this reduces the bias in the data collection [130]. However, over the 3 week period only 201 10-minute blocks of labelled data were collected across 13 people. This is an average of approximately 2.5 hours per person, which is not very indicative of a long term routine. Furthermore, the users could only select from 6 activity labels and 2 of these were never chosen, hence there is little detail available.

Zhu et al. [131] used sticky hierarchical Dirichlet process hidden Markov model

(HDP-HMM), a non-parametric Bayesian model, to segment and cluster raw accelerometer data at different levels of granularity. This approach has the advantage of automatically determining the number of clusters as it is non-parametric. The authors compared this method to LDA and reported that sticky HDP-HMM had lower entropy-based error values for the ground truth segmentations. However, the reported results are for a small dataset, one working day of data for 3 participants with only 5 labelled ground truth labels. Moreover, LDA was applied in a different context to that of the cited work of Huynh et al. [32]; determining activities directly from accelerometer features and location data, as opposed to detecting routines in activity data. Insufficient detail of the application of LDA for this dataset is given and the conclusions drawn are thus very limited.

The studies described here have demonstrated that wearable sensor data used as input to different types of topic models is a promising avenue of research for detecting daily routines. The characteristics of the data must be carefully selected to align with the assumptions of the model, such as exhibiting multiple topics in a document and the target application. Extensions to LDA, such as hierarchical Dirichlet processes have demonstrated the potential to enhance the method by offering additional advantages, such as automatically determining the number of topics. This field of research is still in its infancy and offers scope for expansion in several directions.

3.2.1.3 Environmental sensor data as input

Binary environmental sensor data is investigated by Rogers et al. [132], who apply a hierarchical, iterative approach to using LDA for activity discovery. They evaluate this method on the SCARE corpus, which has been adapted from the original purpose of situated dialogues to produce an event stream. These events related to activities that were carried out in a virtual environment, following a list of tasks and no detailed information of the activities are provided. Although this dataset offers the benefit of having ground truth labels, it is not very realistic in comparison to a free-living real world environment. The results demonstrate that patterns of events in the data at different levels of abstraction can be found. A visualisation of these patterns is presented, however no semantic interpretation is considered and it is not clear how useful these results are for human users.

Rieping et al. [133] considered a different approach by altering the LDA model to combine both the clustering of the sensor data and detection of routines into one model. They tested their novel approach on data collected from PIR sensors, reed and contact switches installed in 5 houses for at least 63 days. They included a preprocessing step that grouped the sensors into 5 key locations and used the number of activations in each location and time of day as features. The results showed that their models outperformed the baseline LDA model on this dataset and they state that meaningful routines are found. However, the interpretation of these routines is not obvious and would not be suitable for end users of a monitoring system. Moreover, by only considering the location of the sensors a lot of detail is lost, for example just because someone enters the bathroom it does not necessarily mean they used the toilet. Whereas if it was known that a flush sensor had been activated this would provide a much higher level of certainty.

A different approach to creating artificial words from binary PIR data is applied by Castanedo et al. [134], similar to that used by Farrahi and Gatica-Perez for global location data. Each word represents a 5 minute sensor activation pattern for a specified sensor, in a given time slot. Documents are then created using the data for one room over one day per document. This method is evaluated on two long-term datasets, 50 and 90 weeks, collected in office environments. Measuring the average perplexity for a 10-fold cross validation suggested that using 100 and 1000 topics on each dataset respectively gave the best fit. The results are interpreted to demonstrate patterns of behaviour in terms of office occupancy. However, the interpretation is very involved and the data is still of a high dimension. Further processing is required to determine useful information from these results. One improvement could be to reduce the size of the vocabulary which contains 38,880 and 83,520 words for each dataset respectively.

Finally, for completeness it should be noted that topic models have also been applied to find activities at a lower level. In particular, Chikhaoui et al [135] used sequential patterns of sensor events as inputs to LDA to find activities. The method was evaluated on 5 datasets, including 2 from the CASAS project, and shown to be a successful method for unsupervised activity recognition. Chen et al. [136] also utilised the CASAS dataset to evaluate their method using a modified LDA model where each document corresponds to only one activity. The documents are created by segmenting the data based on location. LDA is used as an intermediate step by Rai et al. [137] to create features, together with other techniques, that are then used as input to a support vector machine. Raw data is collected from an accelerometer in a smartphone and two different data fusion frameworks are evaluated. Ihianle et al. have considered using a simple topic model, probabilistic latent semantic analysis, on two datasets created from environmental sensors [138, 139] for activity recognition. Moreover, they enhanced this method using domain knowledge extracted from the web.

It has been found that the routines detected by topic models using environmental sensors as input tend to be difficult to interpret and require further processing in order to extract useful information. These results could be improved by including a wider range of environmental sensors rather than focussing on location data generated by movement sensors. Furthermore, the results could be fused with data from other types of sensors to provide additional context and enhance the level of information obtained. Using topic models as an activity recognition method has been shown to be successful. This suggests that a hierarchical approach to topic modelling could be an interesting avenue for research to create a model that can work with raw data and provide results at multiple levels of abstraction.

3.2.2 Other methods applied to the detection of behaviour patterns

3.2.2.1 Detecting daily routines

In addition to using topic models, work has been done to discover daily routines using other methods. For example, Zheng et al. [140] explored the use of collaborative filtering to identify representative routine patterns for individuals even if only sparse mobile phone data is available, such as the location of the user when they make a phone call. As with some of the previous studies using topic models, this work considers daily routines at a high level based on locations of interest, however the technique proposed could be applied to other datasets. The intuition behind collaborative filtering is that there are only a few different behaviour routines people can have and hence given many users' data there will be a large number of other users with similar routines. Details of the algorithms used can be found in [140]. This approach would be most suitable for large, sparse datasets.

The MIT Reality Mining dataset, to which Farrahi & Gatica-Perez and Nabaei et al. applied topic models to discover daily routines, was also analysed by Eagle and Pentland, who collected the original dataset [141]. They used principle component analysis to identify the top characteristic vectors that they term eigenbehaviours [142]. Similarly, the routines found using this approach are at a very high level, such as working late or sleeping in. Eagle et al. consider both routines found for a given individual and also routines associated with an affiliated group of individuals within the dataset, such as business school students. This approach works well for a large, longitudinal dataset where the aim is to characterise behaviours and social interactions at a high level as it provides a large reduction in dimensionality of the data. However, for the purpose of providing data to support healthcare and well-being requirements, too much detail of an individual's activities would be lost.

The Bluetooth proximity data in the MIT Reality Mining dataset was investigated by Azam et al. to determine human behaviour routines [143]. They applied an n-grams technique to identify the longest repeated patterns of proximal Bluetooth devices detected by a specific user. Different levels of variance were considered to account for anomalies in patterns, such as detecting a visitor's Bluetooth device and the frequency of patterns in defined time slots were explored. This work was then extended by also considering a correlation matrix of the repeated patterns discovered to remove sub-patterns that are not unique [144]. However although frequently occurring patterns were highlighted for one user, no interpretation of these patterns is attempted and it is not clear how useful these results are.

A prototype of the University of Virginia smart house monitoring system introduced in section 3.1.3, was installed in a volunteer's residence over 12 weeks [145]. The sensor data was analysed using mixture models, an unsupervised learning method, to detect clusters of sensor firings. A set of performance metrics were used to establish the most significant clusters and it was found that these corresponded to event patterns or routines, such as waking up. The authors suggest that this method can be used to determine a baseline of normal activity for an individual. However, the level of correspondence between the clusters and activities is not clear and has only been investigated for one participant. Additionally, using an unsupervised technique for the discovery of routines directly from raw sensor data offers the advantage of not requiring training data but results in a loss of detail.

Li et al. [146] proposed a method of discovering a flowgraph relating to a person's movement patterns around the home from motion sensor data. Probabilistic deterministic finite automata are learnt, using the Alergia algorithm, to create the flowgraph, from which subflows corresponding to an activity are extracted by applying a weighted kernel k-means algorithm. The identified subflows are mapped to the floor plan of the environment and labelled according to their correspondence with the ground truth. Further analysis using the method introduced by Eagle et al. [142] is used to determine the eigenbehaviours. Li et al. suggest that these correspond to different routines, however interpreting these results to provide useful information is still unintuitive.

Motion patterns were also considered by Yin et al. [147], who took a top-down approach to discovering daily routines and then room-level activities. Rather than directly using the sensor names as done by Li et al., the location of motion sensors e.g. kitchen, are used to model transitions through a living environment. Markov chains are used to model these transitions for each one hour time interval and then merged using a hierarchical clustering approach to find daily routine patterns, such as morning routine. The example Markov chains intuitively correspond to feasible location patterns, however no validation of the method is provided.

A data mining approach was taken by Lin et al. [148], who extracted frequent sequential patterns at the activity and routine level. However, the method appears to rely on the raw sensor data already being labelled. Moreover, it is not shown how the resulting association patterns between activities can be used. It is not clear if this method could be applied to new raw sensor data to estimate what activities and routines are being conducted. Data mining techniques have also been considered as part of the CASAS project. In particular, Rashidi et al. developed a novel frequent sequence miner algorithm, discontinuous varied-order mining, for unsupervised activity recognition that allows for discontinuities and order variation within the sequences [96]. They extended the method to be used with streamed sensor data from real life smart apartments, enabling automatic discovery of activity patterns over time [149]. This provides information for the routine of each individual, which can then be used to look for trends in behaviour. The methods were evaluated and shown to work well on several datasets, both scripted activities and free living [96, 149].

Blanke and Schiele took an alternative view of the challenge to recognise daily routines, based on activity spotting to reduce the amount of data required to classify routines [150]. They demonstrated that a joint boosting framework that uses low-level activity spotters as weak classifiers can reliably discriminate between high level routines. This was evaluated on the UbiComp08 dataset, collected by Huynh et al. [32], showing that using less than 3% of the original data still yielded recall and precision rates over 80% for the routines. This approach is very useful for improving computational efficiency, however it suffers from a loss of data which could reveal important data to clinicians. In particular, it only identifies discriminative features for each routine rather than providing information about the underlying structure.

The methods presented here for detecting daily routines tend to be based on location data and/or have a lack of detail about activities performed. The most comparable method to topic models is the work of Rashidi et al. which discovers activity patterns that correspond to daily routines, whilst allowing for discontinuities and interleaved activities. The patterns are found from raw sensor data and therefore still require additional knowledge of the system architecture and layout in order to interpret the routines. It is not possible to make direct comparisons between these approaches as they have been applied to datasets with very different characteristics.

3.2.2.2 Detecting changes over time

Rather than following a hierarchical approach and looking for routines that are constructed from a series of activities, some researchers have pursued the idea of using the activities found from sensor data to investigate changes over time. This varies from considering overall trends in activity levels to focussing on the patterns for a specific activity of interest, such as using the toilet. This high level overview of changes over time can often be presented in a more accessible form to key stakeholders, such as clinicians and caregivers. However, interesting data about which activities make up a routine may be lost by taking this approach.

Derungs et al. [151] conducted an exploratory analysis of trends in wearable sensor data from stroke patients, from the dataset collected by Seiter et al. [126] described in section 3.2.1.2. The results showed that leg movement and sit to walk duration ratios demonstrated the strongest trends over time. There was still a lot of variability between patients, with both positive and negative trends discovered. It is not clear from this analysis whether the results correspond with clinical observations of progress levels in each patient's recovery and hence how useful they are for clinicians.

The prediction of long term functional health status using environmental sensors, particularly passive infra-red (PIR), has been considered by Robben et al. [152, 153]. Two case studies of elderly participants, living independently, with monitoring systems installed in their homes are presented in [152]. An exploratory analysis of detected location in the house over time and the application of principal component analysis are considered and compared against various health metrics taken every 3 months, used as ground truth. Although these approaches can help to highlight trends in behaviours that can be linked with health outcomes, the results still require interpretation, which is not feasible for clinicians or caregivers. This work is extended in [153] and moves towards addressing this limitation. A regression based approach is utilised in order to automatically select features which are discriminative for different health metrics and represent common concepts for assessing health status e.g. time spent in the living room in the evening. The results show that using features based on changes in measurements yields more accurate results than the static values.

Long term datasets collected at TigerPlace have also been used to investigate behaviour patterns as an early indicator of health decline. One technique proposed was linguistic summarisation, which enables numerical data to be presented using quasi-natural language phrases [154]. This method is based on fuzzy logic calculus of linguistically quantified propositions, such as, 'on some nights the resident had high restlessness'. This methodology allows for uncertainty in the data and provides results in an easily interpretable format for key stakeholders, such as clinicians. Conversely this leads to a reduction in sensitivity to changes and details which may be of clinical importance. A 15 month case study presented for one 80 year old resident demonstrates that contextual knowledge, such as medical procedures undertaken, is very important when analysing the results. This would be difficult to scale up for a large population.

Another approach considered, using the TigerPlace data, is dissimilarity comparison of activity density maps from different time points [155]. Activity density of residents, calculated from the number of sensor firings, can be visualised with colours representing different ranges of activity levels for each hour of the day over a month. This quickly enables general patterns and trends to be seen. A co-occurrence matrix can be used to measure and quantify the dissimilarity between two maps, based on features related to the regularity of the activity data as well as the average motion density and time away from home. The direction of changes is not directly reported however. Results from three case studies showed that observed trends could be linked to health status. Moreover, health professionals found the maps useful to support their job and now use them regularly. Detailed specific knowledge, such as when the cleaner comes, is still required to interpret the maps as this causes a large increase in the number of sensor firings.

More recently, a classification method was used to select features, extracted from motion and bed sensor data collected at TigerPlace, most relevant to capturing early health changes [156]. It was determined that features linked with bathroom visits, sleep patterns and socialising were the most important. These features were used to create an automated health alert system based on comparison of sensor values with a two week moving baseline, personalised for each person. High thresholds were used for the alerts, resulting in approximately 50% false alarms. The clinical relevance of all of the alerts generated were rated by clinicians to provide a ground truth. One domain knowledge based machine learning algorithm and three supervised learning algorithms, with different feature spaces, were compared to automatically determine the relevance of alerts. The results demonstrated the fuzzy pattern tree method, based on domain knowledge performed the best on a dataset of 21 residents for one year. The system enabled the early detection of a variety of health conditions, including urinary tract infections, heart failure, delirium and hypoglycemia. This method works well for TigerPlace residents but may not transfer directly to the wider population.

The idea of an activity curve was introduced by Dawadi et al. [157] as part of the CASAS project. It models the probability distributions of a specified set of activities at different times of day for a whole day. Minor day-to-day variations in the timing of routines are accounted for by aggregating activity curves within a window of a few days. Longer term changes are discovered by comparing the curves at different time points using the proposed algorithm; a permutation-based two-sample test with a distance metric of symmetric Kullback-Liebler divergence. This method was validated on a synthetic dataset with known changes and evaluated on real long term smart home data from 18 participants aged 73 or older. The results demonstrated a statistically significant correlation to the Timed Up and Go Test, a standard clinical test for mobility based health but not the RBANS, a measure of cognitive health. Continuous changes can also be detected and the results indicate that larger variations can be associated with a decline in health [157]. The hierarchical nature of many activities was not considered and only a limited set of labels was utilised. Extending this further could improve the correlations and provide more useful information for healthcare providers.

Four case studies from the assisted living pilot at the University of Virginia were analysed for behavioural patterns [117]. The motion sensor data was used to determine activity level and amount of time spent in each room for each hour of the day. These features enabled estimation of participant's circadian activity rhythms and their deviations. It was demonstrated that this can be used to model behavioural patterns and changes that may be consistent with disease onset. This approach is strongly linked with time of day and therefore may be less effective for populations with a less structured and predictable routine.

Elbert et al. [158] also investigated deviations in circadian rhythms, focusing on people suffering from dementia. Activities of daily living (ADLs) were detected from sensor events using a combination of techniques, including case-based reasoning, rule-based reasoning, time maps and fuzzy logic. Long-term behaviour monitoring is achieved through a combination of an activity score calculator, sleep monitor and circadian rhythm score (CRS) calculator, which are used by a trend analyser to detect critical changes in behaviour. In particular, the CRS calculator is designed to analyse a person's daily routine giving a score from 0 to 1, representing the degree of deviation from normal. This is achieved through a combination of three approaches: cosinor analysis based on the theory of circadian rhythms as a special representative of regression analysis; a histogram-based approach based on movement data and a probabilistic model of behaviour based on a person's ADLs. Details of these methods are given in [158]. This approach appears to be very comprehensive, however it is not clear how much computation power is required and whether it would be feasible to implement such a system outside of an ambient assisted living environment.

Changes in activity levels and patterns over time have been linked with clinical outcomes in a variety of studies. The changes detected do not generally provide specific details about the structure of an individual's routine, rather they tend to indicate variations in overall activity levels. This information can be presented in a usable and useful way to clinicians and caregivers. This level of information may be sufficient for certain applications, but a greater level of detail about which routines and activities are changing, the level and direction of the change could further exploit the data collected, enhancing the benefit of a monitoring system.

3.3 Detecting behaviour patterns in nutrition data

A residential healthcare monitoring system offers huge potential to collect a wealth of data. In addition to detecting activities of daily living from different sensors, nutrition data can be gathered through the use of cameras and logging applications on mobile phones [31, 159, 160]. Good nutrition is fundamental to a healthy lifestyle [161] and many chronic diseases are related to diet [16]. Collecting data on food intake is paramount for investigating associations between diet, health and the occurrence of disease [159]. Eating behaviours and patterns can be directly related to chronic diseases such as obesity, cardiovascular disease, cancer and diabetes [162–165].

This section reviews the state-of-the-art in technology based solutions for monitoring food intake that could be utilised in a residential environment. Background information on the assessment of food intake at different levels is given. In particular, the concept of meal-based analysis is introduced and how this can be used to explore the relationships between diet and health. Finally, the challenge of defining different eating events, especially distinguishing between meals and snacks is considered. Existing definitions used in the literature and qualities affecting perceptions are highlighted. More complex food based definitions are reviewed and the use of machine learning algorithms to identify relevant food groupings is addressed.

3.3.1 Automated nutrition monitoring using sensors in a residential environment

As cameras are already a proposed part of many residential healthcare monitoring solutions, using them to detect food intake is a logical step. He et al. [166] proposed a classification approach based on k-nearest neighbours and vocabulary trees for identifying images of foods consumed during an eating event, taken on a smart phone. A classification accuracy of 64.5% was achieved for the most probable label given to each segmented food item. Ciocca et al. [167] recently presented a benchmark dataset of real canteen trays, containing multiple instances of food items. They developed an automatic tray analysis pipeline and results showed approximately 79% accuracy of food was achieved using convolutional-neural-networks-based features.

The idea of using cameras for dietary assessment was developed further by Dehais et al. [168] who created a novel system to calculate portion sizes using two images from mobile devices. The algorithmic framework presented enables 3D reconstruction and volume estimation of various food items on a plate. The results demonstrate this system is more accurate than existing solutions and user estimates. However, there are currently several limitations due to the requirements for input pictures, type of serving dish and the need for a size reference object in the image. Furthermore, Liu et al. [169] have considered how the use of images taken on mobile devices for dietary assessment can be handled in a real-world scenario, with a particular focus on real time processing and low energy consumption, whilst still achieving state-of-the-art recognition accuracy. This was achieved through the use of deep learning algorithms and the use of cloud based servers for processing.

In addition to detecting food intake in terms of what is consumed, there is also interest in eating behaviours, such as choice of food and number of bites or chews. There are a number of different sensing technologies that are being investigated. The aim is to reliably and unobtrusively gather information on what, how and when items are consumed. The state-of-the-art solutions currently focus on a specific aspect of dietary monitoring, each offering unique advantages but still having severe limitations [170, 171]. As technology and research in this area develops these innovations can be integrated to provide a detailed automatic dietary monitoring system. Some examples of state-of-the-art solutions for monitoring eating behaviour are highlighted, a more detailed review is beyond the scope of this thesis but can be found here [170].

A smart dining table was proposed by Manton et al. [172] to track eating behaviours during a meal. They found that combining a wearable inertial measurement unit with a multi-touch tabletop computer running Microsoft Surface performed better than a kinect for detecting bites and which plate the food was taken from. Zhou et al. [171,173] also designed a smart table surface using a novel smart tablecloth with a fabric based pressure matrix and a dining tray with pressure sensors on the feet. This system enables detection of detailed dining-related activities such as scooping and cutting, the type and weight of items consumed, including drinks. Furthermore, a novel algorithm to detect the weight of individual bites consumed during unrestricted eating has been developed for use with a simple table embedded scale [174].

Table top sensor systems have the limitation of not being easily portable, whereas wearable solutions can be used when eating in different situations, such as on the move or in front of the TV [170]. One wearable solution for detecting eating events and chewing is to integrate electromyography electrodes into custom 3D printed eyeglasses. The glasses can be worn unobtrusively in a free living environment for a full day and can classify between different food types based on textures and hardness [175,176]. Furthermore, the feasibility of fabric based sensors embedded in a shirt collar to detect swallowing has also been demonstrated [177]. Inertial sensors worn on the wrist can be utilised to detect eating events, similar to general activity recognition discussed in section 3.1.4, but developing algorithms to be more specific [178].

3.3.2 Assessing and analysing dietary intake

Monitoring and assessing food and nutrient intake data for individuals is a complex process. There are a variety of different methods employed by scientists, including food diaries, 24-hour dietary recall and food frequency questionnaires. Each method has its own strengths and limitations and the most appropriate method depends on the objectives of a study [179]. Understanding nutrient intake at both a macro and micro level is important for investigating links with health outcomes. Food composition or nutrient databases are used to calculate overall nutrient intake from food consumption data, although these are not standardised and can be a significant source of error, particularly when making comparisons across multi-national datasets [180].

There are many benefits to considering food intake at the individual food item level, but the variety of items means it is difficult to use this level of information to communicate with the public [181]. Similar items are often grouped together into food groups, such as puddings, fruit and cheese. These groups may then be split further into subsidiary food groups, such as sponge puddings (manufactured), citrus fruit not canned and cottage cheese [182]. These groupings are not standardised and can vary in number and content between studies [182, 183]. Only considering food items and the constituent nutrient content does not enable a full analysis of food intake and associations with health outcomes, often leading to conflicting findings [184]. People often eat a variety of food items together, for example a sandwich could consist of wholemeal bread, low-fat spread and cheddar cheese. The interaction between different foods and nutrients consumed together should also be considered [181, 185]. Dietary pattern analysis has been utilised widely amongst the nutritional epidemiology research community and related to chronic diseases [163, 164, 181].

Recently interest in understanding how the composition and patterns in the consumption of meals impact the relationship between diet and disease has increased. A meals-based approach would complement existing dietary advice, which is given at a food or nutrient level [181, 184].There are strong advantages to analysing food intake at a meal level, however one key drawback is the difficulty in comparing results across studies due to the variation in definitions of meals and snacks [17, 18, 184, 186]. In particular, studies that consider the impact of eating frequency on health outcomes, such as obesity are highly inconsistent, reporting a full range of inverse, null and positive associations. This is due to a variety of methodological problems: choice of assessment method; adjustment for confounding factors; under-reporting by certain participants and the lack of consensus about what constitutes a snack, meal or eating event [187, 188]. Similar challenges occur when investigating the impact of skipping breakfast on cognitive performance and appetite control, particularly as the meaning of breakfast is often subjectively interpreted by participants with no clear guidance from researchers [189].

3.3.3 Defining eating events

3.3.3.1 Influence of language and culture on definitions of eating events

It is well understood that there is little consensus within the nutrition research community as to what constitutes a snack [17, 18, 190–193]. Furthermore, the use of different phrases with the root word ‘snack’, e.g. snack food or snacking can cause a large difference in interpretation by individuals [194]. Hess et al. compare the recommendations relating to snacks, snack foods and snacking from dietary guidelines published by countries around the world. The variety of language used demonstrates the diversity of definitions for snacks and the challenges in communicating healthy eating information to the public [191]. Moreove, this is further demonstrated in other languages that have a much greater variety of words to describe eating events outside of meals [193].

A variety of factors can affect the perception of what a snack is, including time of day, type of food, amount of food and level of hunger [17, 192]. Leech et al. [190] compare a variety of definitions for eating events used in the literature. This includes participant defined labels, with five types of meals and four types of snack, including a ‘beverage break’. Other definitions are based on the time of day and energy content of eating events to distinguish them as meals or snacks, with a maximum of three meals possible and all other events deemed as snacks. The remaining six definitions highlighted are neutral, only defining an eating event, using criteria related to time interval separation and minimum energy content, but not distinguishing between meals and snacks. Another definition is based on contribution to the total energy intake (TEI) in a 24-hour period, with those eating events providing greater than 15% of TEI considered a meal and those lower, a snack [195].

Johnson and Anderson propose a qualitative definition for snacks, excluding beverages consumed alone [192]:

> “A snack is composed of solid food(s), including those typically eaten with a utensil (with or without a beverage) that occurs between habitual meal occasions for the individual, is not a substitute for a meal, and provides substantially fewer calories than would be consumed in a typical meal.”

This definition is based on identified characteristics from the literature, but it is offered for initial consideration and acknowledged that a collaborative effort among all relevant stakeholders is required [192]. Warde and Yates provide very insightful analyses into types of eating events based on a survey of eating patterns conducted in Great Britain in 2012, focussing both on snacks [193] and meals [196], including a comparison with how meal content has changed since the 1950s. They suggest that the definition for a snack would have the following qualities [193]:

> "Small portions of portable foods, easily prepared, and consumed casually at simple and quick events."

In addition, a more specific definition has been proposed for breakfast by O'Neil et al. [197]:

> "Breakfast is the first meal of the day that breaks the fast after the longest period of sleep and is consumed within 2 to 3 hours of waking; it is comprised of food or beverage from at least one food group, and may be consumed at any location."

Further guidance is also provided with regards to criteria for a quality breakfast in terms of nutrient and energy balance. St-Onge et al. suggest combining this definition, with the definition based on TEI, to cover all eating events without referring to the time of day, as it excludes subgroups of the population, such as shift workers [195].

Situational and environmental cues, such as eating alone or with others, can also influence perception of eating event classification [193, 196, 198]. Young children, aged 4 to 6 years, are already learning associations between eating cues and the classification of eating events. These eating habits can endure over time and therefore influencing children's perceptions from a young age could help to encourage a healthy and well-balanced diet [199]. Younginer et al. focussed specifically on snacks for children from the perspective of low-income caregivers. They highlighted five dimensions that influence the definition of a snack, arising from semi-structured interviews: type of items consumed, portion size, purpose of eating event, location and time, both time of day and time for preparation and consumption. These resulted in a proposed definition for a child's snack [200]:

> "A small portion of food that is given in-between meals, frequently with an intention of reducing or preventing hunger until the next mealtime."

3.3.3.2 Food based definitions of eating events

Beyond the use of participant defined labels, simple criteria, based on time of day and energy content or qualitative descriptions, some studies have considered how foods are grouped together in eating events. In 1999, Lennernäs and Andersson [201] proposed a categorisation strategy for eating events based on combinations of food categories, known as 'food-based classification of eating episodes' (FBCE). Rather than using specified thresholds to define events, this strategy first categorises different types of eating events and then considers the mean energy and nutrient content. Seven food categories are given based on their nutritional similarity, such as 'high fat density'. A table of criteria on how these categories are combined is used to determine the type of eating event, from a list of four meals and four snacks. These are given labels, from 'complete meal', through to 'no energy snack', with examples provided for each. However, this classification strategy has not been widely adopted by the community [192], perhaps because it does not consider the other qualities and factors associated with eating events.

More recently, a food based classification method with detailed criteria for identifying eating events as meals, snacks or drinks was developed to investigate the association between restrained eating, total energy intake and obesity [202]. All food groups recorded in the UK National Diet and Nutrition Survey 2000, excluding supplements, were allocated to meal, snack and drink lists, based on information from the literature. Eating events were considered to be all items consumed within every 60 minute period, starting from the time of the first record on each day for every participant. The following criteria were then applied to label an eating event as a meal, snack or drink:

Meal All items from meal list OR
More than one item and at least one item from meal list EXCEPT
Two items only, one each from meal and snack lists (e.g. bread and butter)

Snack All items from snack list OR
Two items only, one each from meal and snack lists

Drink All items from drink list OR
Two items only, one each from drink and snack lists (e.g. sugar in coffee)

This method allows for common combinations of food items, which are listed separately in databases, to be categorised in the relevant group. For example, an eating event which only contains a coffee will still be labelled as a drink even if there is sugar in the coffee. However, this method still has limitations, for example a cup of tea where tea, milk and sugar are listed separately would not meet the criteria for the drink category. This highlights the complexity of defining eating events and the influence of underlying coding strategies used in databases on developing suitable criteria.

The concept of generic meals was introduced by Woolhead et al. [181], based on common food group combinations that occurred in the Irish National Adult Nutrition Survey (NANS). The food items recorded in the survey were allocated to 20 food groups and the recorded eating event labels were simplified to five types: breakfast, light meals, main meals, snacks and beverages. Within each type, the most commonly consumed food group combinations were identified, using a data-mining method based on the concept of 'frequent item sets'. These food group combinations were considered to represent generic meals, such as 'cereals and milk and juice' and each assigned a unique code. A total of 63 generic meals were found across the full dataset, excluding all supplement items.

Generally dietary patterns are used in nutrition research to identify trends in overall dietary intake. However, this approach was used at the meal level for a dataset of food intake for Brazilian adults [203]. Three types of breakfast were identified: healthy, traditional and snack. There were five types of lunch: traditional, salad, sweetened juice, western and meats. Finally, for dinner four patterns were found: coffee with milk and bread, transition, traditional and soup and fruits. These groupings of foods could be useful for the specific population studied, however the current naming conventions limit their use in different settings as they rely on assumptions e.g what constitutes traditional and western.

Johnson et al. [204] investigated how foods are grouped based on individuals' hedonic ratings of foods and the relationship to dietary intake. Principle components analysis (PCA) was used to determine the component structures of the food ratings. For each component researchers reviewed the images and selected the following names to describe the groups: Energy dense Main Courses, Fruits, Meats, Desserts, Light Main Courses, Seafood, and Grains. These were compared to traditional food groups, revealing that some groups were an exact match, e.g. fruits, whereas others e.g. energy dense main courses had a poor agreement due to a more complex mixture of foods. Notably, vegetables were not found to be a discernible component in the PCA. Although this does not specifically focus on defining eating events, understanding how individuals group foods in terms of appeal can help dietary guidelines be tailored to have a greater impact on encouraging healthy eating.

As far as the author is aware, at the time of writing, machine learning algorithms have not been directly utilised to investigate how foods are grouped together in eating events. However, there has been interest in the use of machine learning to investigate eating patterns at a meal level, as opposed to a food item level. Hearty and Gibney [185] developed a 2-tiered coding system to represent types of meals based on the foods consumed together. These codes were used as the input to supervised machine learning algorithms to predict a Healthy Eating Index score. Khanna et al. [16] implemented kernel k-means clustering to identify temporal dietary patterns in terms of energy consumed. Clustering was also adopted by Riou et al. [205] to identify temporal patterns in meals for the population of Paris.

Participant defined eating events, definitions based on simple criteria relating to time of day and energy content or qualitative descriptions are inherently biased by language and cultural influences. Some studies have utilised food based definitions, which are linked with ideas on how foods are combined together in meals, rather than relying on time or energy content. This reduces some restrictive assumptions, such as snacks have a lower energy content than a meal. However, these definitions are still subjective in nature due to the choice of categories for grouping foods. The lack of consistency between definitions used across studies poses a huge challenge when analysing eating events to understand how different combinations of foods are linked with diet quality and health outcomes. Interdisciplinary collaborations and advances in machine learning techniques have enabled the use of new methods to be applied to nutrition research. A data driven approach to understanding how foods are combined could help to limit researcher bias and provide a standardised approach for analysing eating events.

3.4 Chapter summary

This chapter has reviewed the literature related to the application of data fusion for healthcare monitoring in residential environments. The rich, complex nature of datasets that can be collected by a monitoring system consisting of a variety of sensors has been highlighted. Understanding these datasets and obtaining useful information for the end users was considered within the context of data fusion frameworks. Research at the low and middle levels of data fusion, in particular activity recognition has received a lot of attention and is a mature field of work. A variety of the techniques and algorithms employed were presented, along with a discussion of the advantages and limitations.

High level data fusion for detecting behaviour patterns has received a lot less attention in comparison to activity recognition and was identified as an important area for further research. Clinicians and caregivers have limited time available to review data from a monitoring system on a regular basis and therefore summarising information at a high level can improve the overall usability. At the same time, it is important not to disregard the detailed data at a lower level as this may need to be referred to for specific situations. Topic models were identified as a promising method for detecting daily routines in activity data based on the work by Huynh et al. [32].

Topic models offer the advantage of finding the latent structure of routines based on a series of activities. Therefore a high level summary is achieved at the same time as retaining knowledge of the underlying details. Furthermore, topic models assume that activities can occur in any routine, as opposed to clustering algorithms which restrict each activity to be assigned to only one routine, which is not reflective of the real world nature of activities. The probabilistic nature of topic models is also beneficial as it represents and helps to quantify the uncertainty that is inherent in complex, real data.

Extensions to the original topic model, LDA, have been implemented in a handful of studies related to activity and routine monitoring. There has also been interest in long term monitoring and detection of changes in behaviour patterns over time. This information can aid early diagnosis and management of chronic diseases, improving quality of life for patients. However, many techniques currently applied to this challenge tend to focus on variations in overall activity levels and do not include data about the underlying structure of routines. Dynamic topic models (DTMs) have been identified as a potential method to address this gap in the research. DTMs can model changes in the structure of routines over time, in terms of the probability of the constituent activities. This can reveal more detailed information about how a person's behaviour varies with time.

In addition to activity data, a residential monitoring system can also be employed to collect data related to other aspects of a person's health and well-being. In particular, nutrition is highly correlated with a variety of health outcomes and eating behaviours can be directly linked with chronic diseases. Therefore the application of topic models to nutrition data was highlighted as a key research opportunity. In collaboration with a nutrition expert from the University of Bristol it was identified that a large research challenge in the nutrition community is the lack of a clear definition of different eating events, such as a snack. This affects the analysis of the impact of different eating behaviours on health outcomes and makes comparisons of results across studies difficult. Topic models can provide a data driven approach to understanding how different food groups are combined together in eating events.

Chapter 4

Detecting Routines in Activity Data

An individual's physical activity levels and ability to perform activities of daily living are key indicators of their health and well-being [3, 15, 206]. Monitoring these can enable older adults and people with chronic conditions to maintain their independence living at home for longer [23]. Moreover, continuous monitoring of activities in a residential environment is important for early detection of disorders and deterioration in health [3, 4]. Many machine learning algorithms have been applied to the task of activity recognition and to a lesser extent, detecting behaviour patterns. Topic models have been identified as showing great potential for detecting routines in activity data, offering the advantages of discovering the underlying structure, allowing activities to occur in multiple routines and quantifying uncertainty in the data.

Several studies globally have collected datasets [62, 89, 90, 94, 97, 113, 118, 119, 141, 158, 206–210] relating to human activities and daily routines. However, the nature and availability of these datasets varies greatly. It is recognised that there is an important need for realistic datasets to be publicly available, unfortunately there can be many restrictions that mean this is not always possible. In particular, such datasets are expensive and difficult to collect; furthermore there can be ethical constraints on releasing data into the public domain. Publicly available datasets allow researchers to both validate other's work and build on it, including making direct comparisons between the performances of different algorithms [206, 208].

This chapter describes the implementation of Latent Dirichlet Allocation (LDA) to replicate the work of Huynh et al. using the UbiComp 08 dataset [32] and the results obtained are compared with those originally published. Following on from the analysis of this previous work, the requirements and collection of a new dataset of daily activities, using a custom built smart phone app, are presented. The results of applying topic models to this novel dataset are analysed and discussed.

4.1 UbiComp 08 dataset

This section investigates the open UbiComp 08 dataset presented in Huynh et al. [32]. The aim is to verify the results given, in particular the application of topic models to activity data in order to find higher level routines. Routines are considered to be a group of activities that often occur together, in varying orders and proportions and tend to last a longer period of time than individual activities. For example, the routine 'office work' will mostly involve the activity of 'sitting/desk activities' but may also include activities such as 'using the toilet' and 'standing /talking'. The UbiComp 08 dataset contains data from one person, wearing two accelerometers, one on their right wrist and one in their right hand hip pocket, whilst carrying out their normal activities of daily life. The data was collected over a period of sixteen days, during waking hours. The sensors sampled at 100 Hz and this data was preprocessed on the device using a sliding window of 0.4 seconds to give mean and variance features at 2.5 Hz. The user also annotated their daily routines and the detailed activities that made up these routines, throughout the collection period. These annotations were recorded using a combination of methods, including experience sampling, a time diary and camera snapshots. After accounting for hardware failures and missing annotations, the final dataset used for experiments included 84 hours across 7 non consecutive days.

The data were analysed using a hierarchical approach with both supervised and unsupervised methods. Firstly, the raw sensor data was classified into activities using the mean and variance features, frequency features, and a time stamp. Initially, three supervised classification methods were evaluated, support vector machines, hidden Markov models and Naïve Bayes, using the annotations for training. Huynh et al. found that the best method for this dataset was a Naïve Bayes classifier and adding the frequency features did not improve the results. The classifier results were then used as the input to a topic model by creating documents of 30 minutes of activity data using a sliding window, shifted by 2.5 minutes each time. Secondly, it was then demonstrated that the whole process could be unsupervised by using k-means clustering to generate discrete labels from the accelerometer data. These labels can then be used as the vocabulary for the topic model [32].

The results showed that topic models can identify topics that correlate well with the ground truth daily routine labels given by the user, both using the output from supervised and unsupervised classifiers. Supervised techniques offer the advantage that the topics found are immediately human-readable because they are constructed from meaningful labels, such as walking. On the other hand, unsupervised methods generate topics where the contents is a collection of cluster labels with no direct meaning. However, unsupervised methods have the significant advantage of not requiring an annotated dataset, which requires substantial effort and is not always accurate [32].

4.1.1 Applying LDA to activity data in UbiComp 08 dataset

Huynh et al. used David Blei's implementation of variational inference for LDA, written in C [51], hence this was chosen in order to replicate their results as closely as possible. The C code was compiled using Microsoft Visual Studio and the executable was called from a wrapper function in MATLAB to allow easy manipulation of both the input data and the results. The program has two main functions which can be called. The first one, 'lda est' estimates a topic model which has the best fit to the data i.e. the Expectation Maximisation (EM) algorithm described in section 2.2.2 is executed to determine the latent variables at the document and corpus level. An option is selected to define whether the topic distributions should be initialised randomly or smoothed from a randomly chosen document. An initial value for α, the Dirichlet hyperparameter for the per document topic proportions, and the number of topics K must also be specified.

There is a settings file, which can be altered as required. This states whether to make the initial value of α fixed or estimated as the model is learnt. It also contains settings to define the convergence criteria and maximum number of iterations that should be executed for the variational EM algorithm and coordinate ascent variational inference. The second function, 'lda inf' performs variational inference on new data using a previously estimated LDA model. Again, it uses a settings file to configure the convergence criteria and maximum number of iterations.

In order to use activity data as an input to a topic model it needs to be split into the equivalent of documents that make up a corpus. The same approach as taken by Huynh et al. was used to make the results comparable. The full list of all activities is used as the vocabulary, where an activity is the equivalent of a word. To create documents, a sliding window was applied to the dataset, where each window is a time slice of data that is equivalent to the words in a document. This means that each time slice contains a mixture of routines over a period of time. The routines to be discovered from the activity data are the equivalent of the topics in documents. The

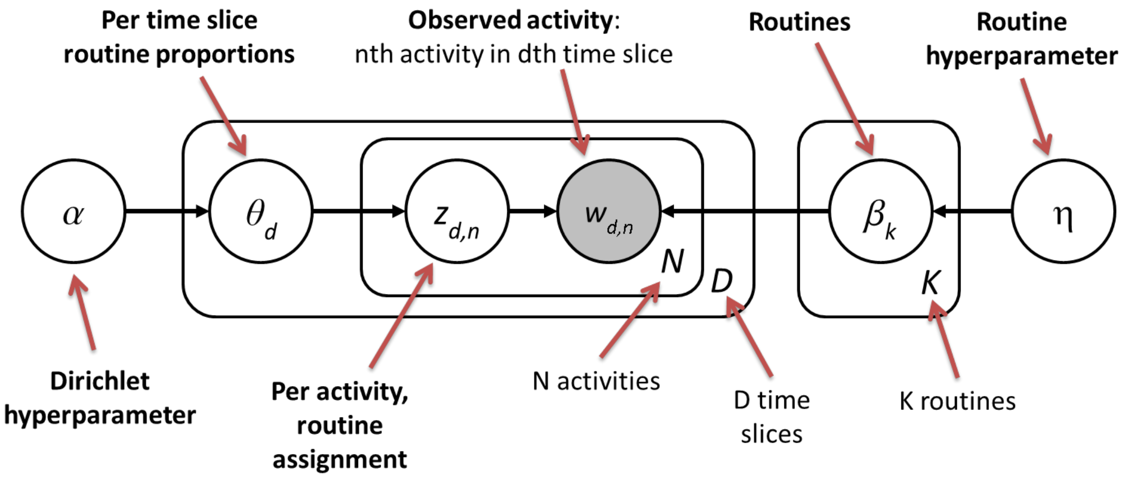


Figure 4.1: Applying LDA model to activity data, see figure 2.6 for original application.

total dataset is equivalent of a corpus of documents. The representation of the LDA model, in the context of activity data for discovering routines is given in figure 4.1.

A window length of 30 minutes, shifted by 2.5 minutes at a time, was used as Huynh et al. [32] found that the performance of the topic models decreased when using windows of a smaller size. The number of routines was set to 10 and the Dirichlet prior α was 0.01. These settings were also used for this work to allow direct comparison. The default settings for the maximum number of iterations and convergence criteria were used as these were not specified by Huynh et al. Furthermore, it was chosen to set α to be estimated as part of the model so that the best fit to the data was found and a random initialisation was used for the routine distributions. There were 37 activities in the vocabulary and each was assigned a unique numerical label.

The LDA C code uses the assumption that the words of each document are exchangeable to be able to represent each document as a sparse vector of word counts. Data is read from a file, where each line represents a document and is of the form:

[N] [word1label]:[count1] [word2label]:[count2] ... [wordNlabel]:[countN]

where N is the number of unique words in the document, word*n*label is the numerical label representing the *nth* unique word and count*n* is the number of times that word appears in the document. A data file of the correct format was generated at the same time as splitting the dataset into time slices by applying a histogram function to the labels in each time slice and printing the relevant information to the file. This was done for 6 days in the dataset and used to estimate the model.

A file of the same format was generated for the remaining day in the dataset so that inference could be performed on this data to predict the routines for that day.

The ground truth labels provided in the dataset for this day were also collected in sliding windows, the same way that the documents were generated. The mode of the labels was used to represent the routine for each period of time.

4.1.2 Routines detected in UbiComp 08 dataset

The results of the estimated model for 6 days of data are output to a file where each line represents a routine. It lists the log probability of the activity given the routine, for each activity in the vocabulary. This file was analysed to find the top activities, with a probability greater than 0.02 [32], in each routine. These were listed in descending order of probability, as shown at the bottom of figure 4.2.

These routines were analysed qualitatively and subjectively assigned a mapping between the routine number and the most appropriate choice from the list of ground truth labels, based on the most common activities that occur in the routine. The ground truth labels in the UbiComp 08 dataset are: ‘Lunch’, ‘Dinner’, ‘Office work’, ‘Commuting’ and ‘Unlabelled’ to cover everything else. For example, routine 1 has ‘sitting desk activities’ with a probability of 1.00 (to the nearest 2 decimal places) and hence is obviously equivalent to the ‘office work’ routine. However, routine 7 is more ambiguous as it includes ‘driving car’ with a probability of 0.43, which is likely to be part of the ‘commuting’ routine and ‘discussing at whiteboard’ with a probability of 0.35, which is likely to be part of the ‘office work’ routine. However, it was decided to map this routine to the ‘commuting’ routine ground truth label as the associated activity had a higher probability. The mappings chosen for each routine are given in table 4.1.

Table 4.1: Mapping of predicted routines to ground truth labels for UbiComp 08 dataset

Routine #	**Ground Truth Mapping**
1	Office work
2	Unlabelled
3	Unlabelled
4	Lunch
5	Unlabelled
6	Unlabelled
7	Commuting
8	Dinner
9	Unlabelled
10	Commuting

Inference was performed on the data for the remaining day, generating the free variational Dirichlet parameters, γ, as defined in section 2.2.2. Subtracting the prior Dirichlet parameters, α from these posterior Dirichlet parameters approximately yields the expected number of activities allocated to each routine for a particular time slice. This was used to determine the probability of each routine for a time slice and hence the routine with the highest probability was considered the predicted routine for that time slice.

Using the mappings, a graph of the ground truth label and predicted routine for each time slice in the remaining day was plotted, as shown at the top of figure 4.2. It can be seen from this that the predicted routines have a good match to the ground truth. There are some small differences, for example it was predicted that the lunch routine was slightly shorter than it was recorded to be. Additionally, a short period of the 'unlabelled' routine was predicted both before and after the commuting routine, which was not labelled in the ground truth. However, ground truth labelling itself is very subjective and although the subject may not have deemed it necessary to label these periods of time as separate routines it does not necessarily mean that it is invalid to do so.

In addition to comparing the predicted routine with the ground truth the results were also visualised by plotting the probability of all routines for each time slice, as shown in the second graph in figure 4.2. This helps to identify why discrepancies occur between the predicted routines and the ground truth. The graph shows that during the main part of the day when the subject was doing office work there is one routine which is very dominant, with a probability of nearly 1 for the majority of the time. However, around the period that relates to the end of the working day, commuting and dinner it can be seen that many routines are activated, all with similar probabilities. This demonstrates that there is a lot more uncertainty as to the prediction of what is happening during this time. However, these results can still be interesting as the uncertainty is quantified by the probabilities.

The original results published by Huynh et al. [32] are included in figure 4.3 for comparison. It can be seen that although the routines found by each model are not exactly the same there are some strong similarities. For example, both models have a routine containing the activity 'sitting desk activities' with probability 1. Furthermore, both models have routine(s) that can be mapped to each of the ground truth labels, although Huynh et al. do not explicitly perform this mapping or plot the predicted routines. Moreover, in the original model the routine of 'office work' is shared between several predicted routines, whereas in the replicated results it is dominated by just one predicted routine, with more predicted routines relating to the 'unlabelled' routine.

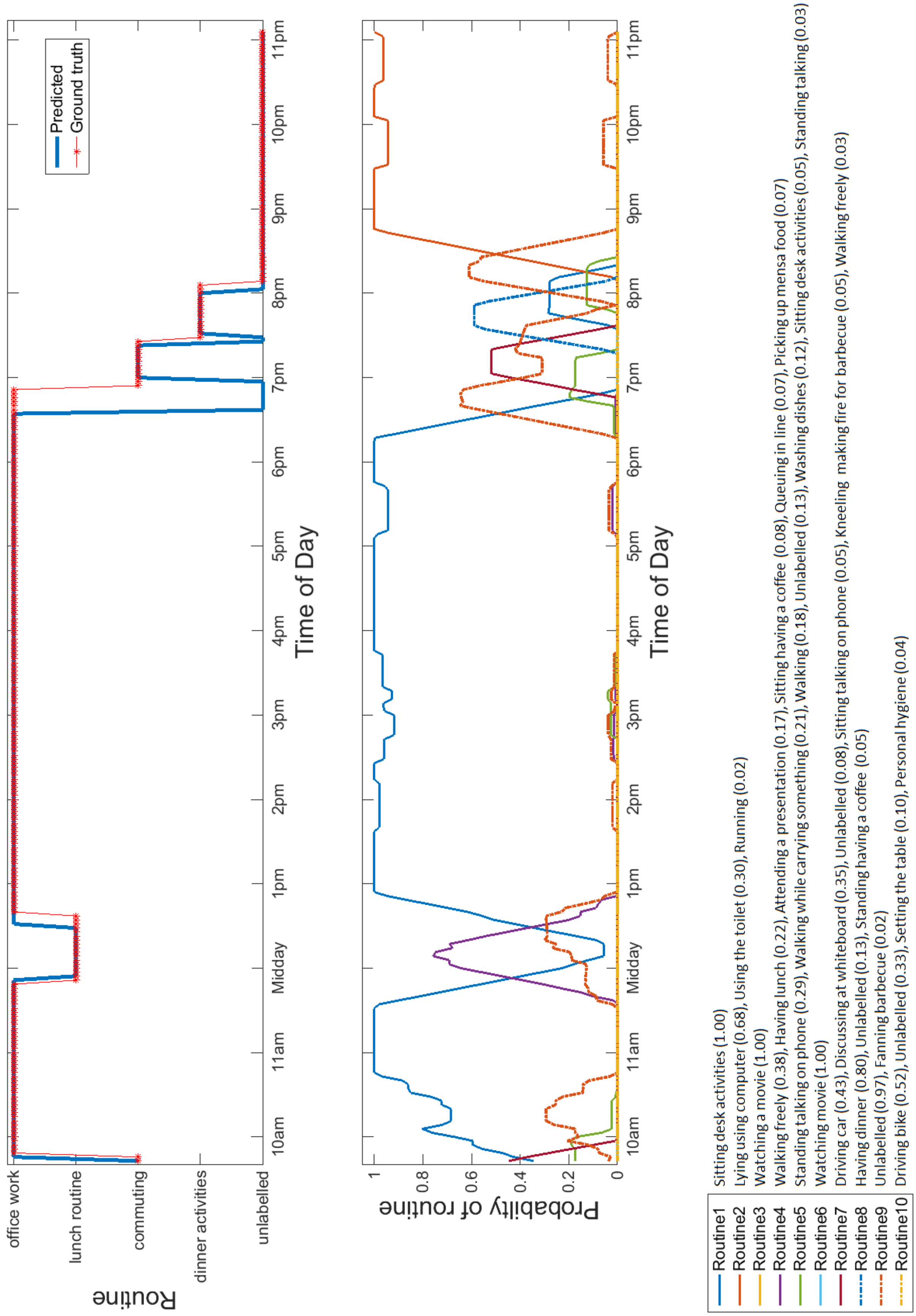


Figure 4.2: Replicated results of topic model for UbiComp 08 Dataset

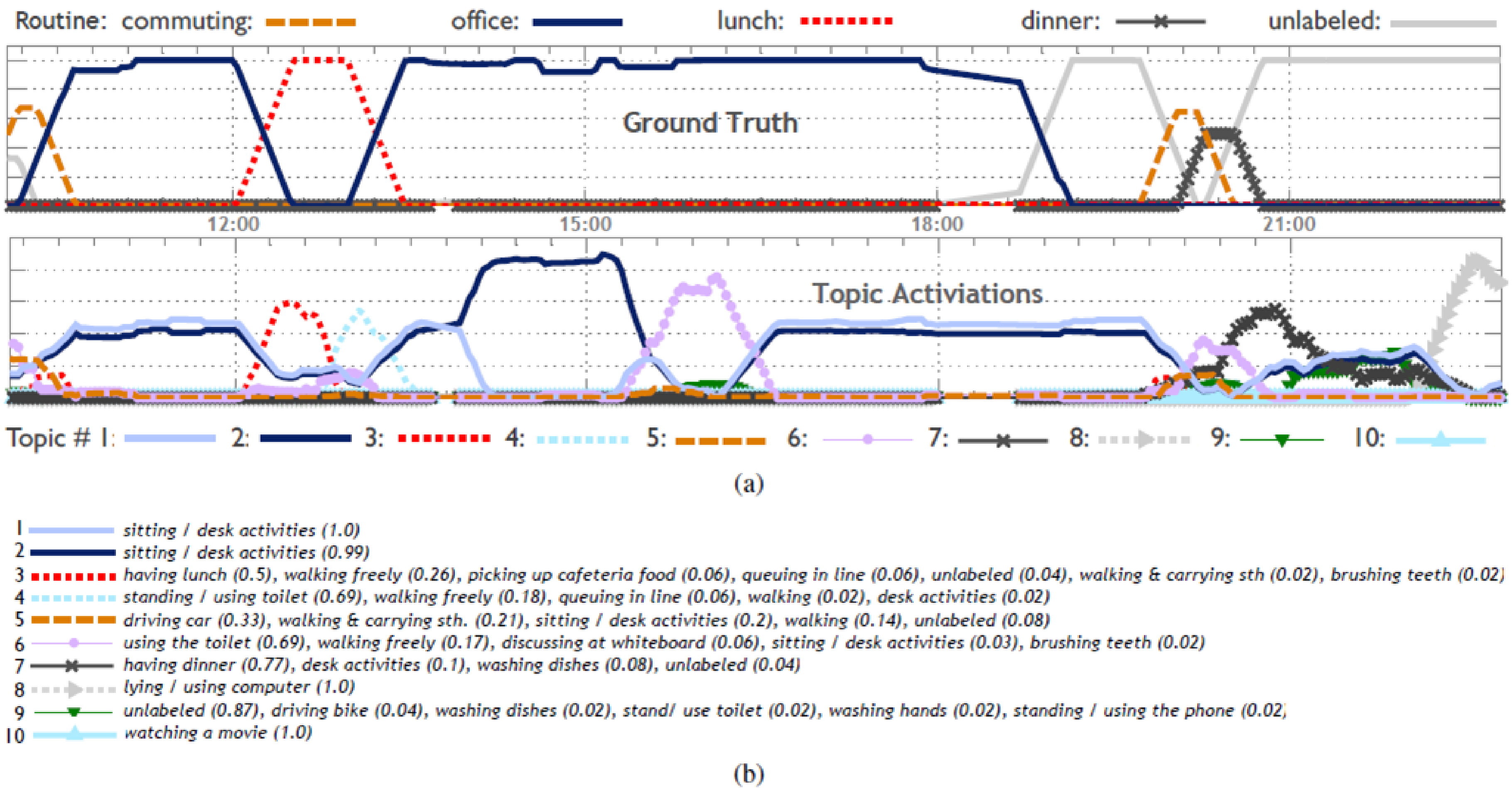


Figure 4. (a) Ground truth and topic activations for one day, based on a vocabulary of learned activity labels. (b) Contents of the ten estimated topics. The numbers in brackets indicate $p(w|z)$, i.e. the probability of the activity label w given the current topic z (labels w with $p(w|z) < 0.02$ are not shown). The distributions were estimated from six days of data. (a) shows the inferred topic activations for the day that was left out during training.

Figure 4.3: Original Huynh et al. results of topic model for UbiComp 08 dataset. Figure used with permission from authors [32].

Table 4.2: Accuracy of predicted routine compared with ground truth labels

Day1	**Day2**	**Day3**	**Day4**	**Day5**	**Day6**	**Day7**	**Average**
80.1%	68.2%	78.7%	84.0%	93.4%	73.8%	95.0%	81.9%

A leave-one-day-out cross validation was conducted using the same settings as previously. For each day a topic model was estimated using the other six days of data and then inference was performed for the left out day. The discovered routines were mapped to the ground truth labels using the same process as before. Using these mappings the most likely predicted routine for each time slice was calculated. These mapped predicted routines were directly compared to the mode of the ground truth labels for each time slice. The accuracy was calculated as the number of time slices with matching labels divided by the total number of time slices. The results, given in table 4.2, show that there is a good match between the mapped predicted routines and the ground truth labels across all days.

Overall, this section has confirmed the results of Huynh et al. and demonstrated that LDA can successfully detect patterns of activities that reflect a person's routines. It has been demonstrated that even when the same dataset is used the random initialisation of the topic distributions can lead to different results with regards to the exact proportions of the activities in the detected routines. However, the results still show strong correlation with the ground truth labels and can be used to explore the high level structure of daily activities. The ability of topic models to quantify uncertainty is a key property as it indicates a level of confidence in the results at any time and reflects the ambiguous nature of human behaviour, such as concurrent and overlapping activities.

At the time at which the work for this thesis was conducted, as far as the author was aware, the work of Huynh et al. was the only published work applying LDA to a publicly available dataset of real world activities with ground truth routine labels to discover routines. A method may provide good results for a particular dataset but this does not guarantee that it will perform as well on different datasets. Applying LDA to another dataset with similar properties helps to validate the method and demonstrate that it can be generalised beyond a specific dataset. The next section considers the collection of a novel dataset of daily activities and associated routine labels which can be used to validate the use of LDA to discover activity patterns reflective of daily routines.

4.2 Collection of a novel daily activity dataset

This section highlights the properties of an ideal dataset for daily routine detection from activity data collected in a residential environment. The details of a mobile phone application developed for collecting activity labels and ground truth routines are presented. The application of LDA to this new dataset is described and the results are analysed and discussed. The challenges involved in data collection are discussed and improvements made to the logging application are given.

4.2.1 Dataset requirements for research into behaviour patterns for healthcare monitoring

There are few published results of applying topic models to daily activity data. At the time this work was conducted the only known published results were using the UbiComp 08 dataset from Huynh et al. [32]. Further work has since been conducted by Seiter et al. [125] to compare results with other datasets, one of which, the Opportunity dataset [113], also has daily activities. However, the Opportunity dataset, although very detailed, was collected in a simulated environment within a laboratory, over a short time period for each participant, approximately 20 minutes per run through the script. Therefore it is not as applicable for the task of daily routine detection of activity data in a residential environment. The UbiComp 08 dataset was provided with a vocabulary of 37 activities, with one being ‘unlabelled’ and two activities that did not occur during the seven days. This size of vocabulary was sufficient to provide interesting results that demonstrate the potential of topic models for identifying routines within a user’s daily activities.

Seiter et al. explored the robustness of topic models by developing simulated datasets based on the UbiComp 08 dataset. They found that the key dataset properties for providing stable topic model performance were the duration of routines, the amount of data and the specificity of routines [124]. In particular, it was reported that routine duration must considerably exceed, by a factor of more than 2, the document length, which was set to 30 minutes and is also influenced by the total amount of data available. Less than 5 days of data will have a large effect on the stability of the topic model and it is recommended that at least 14 days of data is used for good performance, particularly for infrequent routines. Furthermore, routines need to be specific and hence the activities that make up a routine should not overlap with other routines by more than 5% [124].

Based on this and the aims of the SPHERE project, a reasonable dataset would have the following properties:

- Participant(s) from target cohort
- Data collected for 14 + consecutive days for 24 hrs/day
- Data based on body worn sensors, environmental sensors and RGBd cameras
- Specific, labelled routines with less than 5% overlap
- Collected in a residential environment

4.2.2 Development of mobile phone application for logging activities

Unfortunately, it is infeasible to collect a dataset with the properties listed from raw sensors within the time-scale of this thesis as the deployment of the SPHERE platform is due to occur in years 4 and 5 of the project. It was decided to collect a dataset that provided a ‘gold standard’ representation of the first stage. In other words, for the user to directly label the activities they are performing as if they had been automatically generated by a classifier, such as Naïve Bayes, from the raw sensor data. Huynh et al. [32] demonstrated that raw sensor data can be classified into activities using mean and variance features from the accelerometer data and a time stamp. However, a classifier does not have 100% accuracy in activity recognition and therefore a dataset generated directly from user annotation and assumes ‘perfect labels’ has the limitation that it does not represent the noise that occurs in a dataset generated by a classifier.

A mobile phone application was developed in order to log daily activity data and the associated ground truth routine labels. This was developed for the *Windows Phone 7.8* platform as this hardware was already available for use. The aim of the application was to make it easy to record the time a new activity was started and which routine this activity was associated with. It was decided to pre-set the labels for the different routine categories so that they were consistent and suitable for use as ground truth labels. These were based on the work by Huynh et al. [32] and extended to provide more detail for non-work related routines and to include a multi-routine option for activities, such as ‘using the toilet’ that occur as part of several different routines. The final list of categories is: wake up, travel, work, lunch, dinner, relaxation, exercise, multi and other.

The app was designed to give the user flexibility over the labels used for activities so that it could be customised for each user’s needs. The user could select the ‘add new activity’ option from the menu at the bottom of the app, as shown in figure

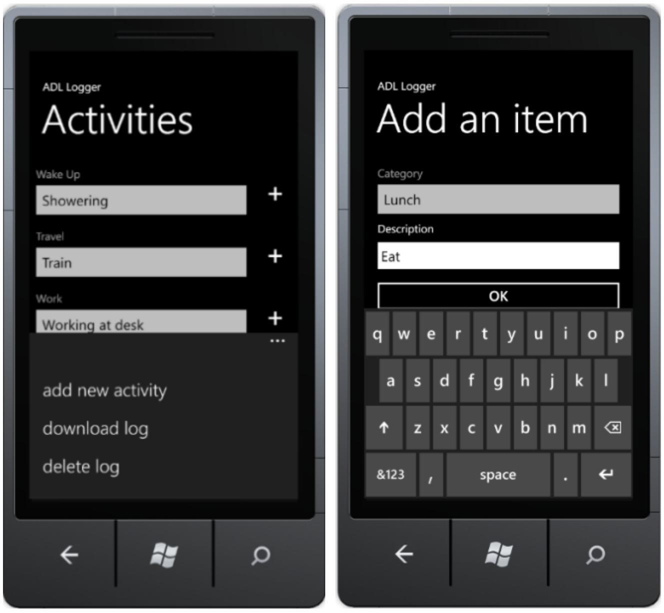


Figure 4.4: ADL Logger App (a) Menu options (b) Add new activity screen

4.4(a), which brings up the screen shown in figure 4.4(b). The user can then select the routine category that the new activity fits into from a drop down list of the pre-set options and then type any description they want for the activity itself. The new activity is then saved as part of that user's vocabulary and will be available for selection on the main screen in the future.

The aim was to make the app quick to use, so as not to be too intrusive in the user's daily life. To this end, to log an activity the user simply has to open the app, select the relevant activity from the drop down list for the corresponding routine and

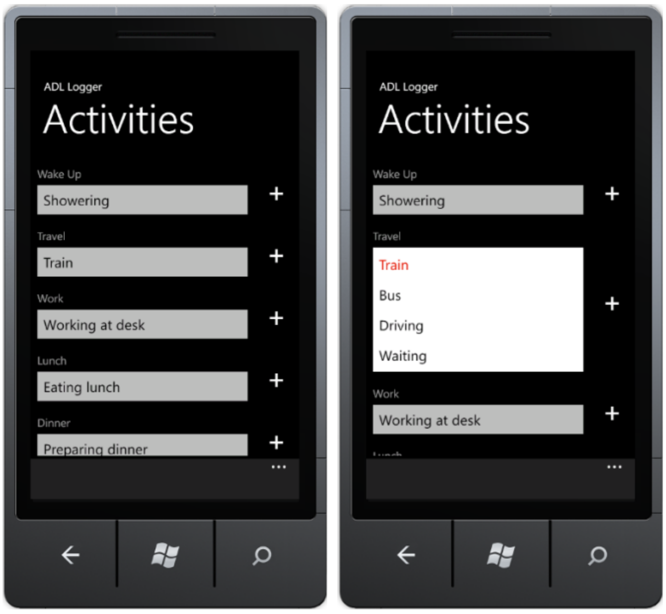


Figure 4.5: ADL Logger App (a) Main screen (b) Selecting an activity

press the '+' button next to it, see figure 4.5(a). If the activity to be logged is already selected then the user only needs to press the '+' button, making the process quicker. If there are only a few options in a routine then the drop down list will appear in situ, see figure 4.5(b) and if the list is longer it will open on a separate screen so it is easier to see all of the items and find the correct one as efficiently as possible.

Vocabulary

Activity (A)	Routine (R)
1 Showering	1 Wake Up
2 Personal hygiene	1 Wake Up
3 Eating breakfast	1 Wake Up
4 Eating snack	8 Multi
5 Drinking	8 Multi
6 Working at desk	3 Work
7 Preparing drink	8 Multi
8 Attending meeting	3 Work
9 Driving	2 Travel
10 Bus	2 Travel
11 Train	2 Travel
12 Waiting	2 Travel
13 Other	9 Other
14 Preparing breakfast	1 Wake Up
15 Dressing	1 Wake Up
16 Gym	6 Exercise
17 Climbing (indoor)	6 Exercise
18 Phone call	8 Multi
19 Using toilet	8 Multi
20 Walking	8 Multi
21 Watching television	7 Relaxation
22 Sitting talking	7 Relaxation
23 Preparing lunch	4 Lunch
24 Eating lunch	4 Lunch
25 Preparing dinner	5 Dinner
26 Eating dinner	5 Dinner
27 Ironing	9 Other
28 Ironing	9 Other
29 Shopping	9 Other
30 Sitting relaxing	7 Relaxation
31 Taking medication	9 Other
32 Sleeping	9 Other
33 Personal hygiene	8 Multi
34 Queuing	4 Lunch
35 Standing talking	9 Other
36 Packing bag	1 Wake Up
37 Giving presentation	3 Work
38 Doing chores	9 Other
39 Bathing	9 Other
40 Attending presentation	3 Work

Log

R	A	Date	Time
9	13	31 7 2014	07 46 58
8	19	31 7 2014	08 04 15
1	1	31 7 2014	08 07 17
9	13	31 7 2014	08 18 52
1	15	31 7 2014	08 25 56
9	13	31 7 2014	08 34 59
1	14	31 7 2014	08 40 29
1	3	31 7 2014	08 42 15
1	36	31 7 2014	08 47 29
9	13	31 7 2014	08 53 23
2	9	31 7 2014	08 56 41
8	20	31 7 2014	09 59 03
3	8	31 7 2014	10 08 57
8	19	31 7 2014	10 35 39
8	20	31 7 2014	10 38 10
9	13	31 7 2014	10 44 34
3	37	31 7 2014	11 03 33
3	40	31 7 2014	11 24 13
3	37	31 7 2014	11 46 35
9	35	31 7 2014	12 01 24
8	20	31 7 2014	12 10 29
3	8	31 7 2014	12 16 40
4	24	31 7 2014	12 26 30
3	8	31 7 2014	13 05 43
8	20	31 7 2014	15 35 57
7	22	31 7 2014	15 42 36
8	19	31 7 2014	16 06 40
8	33	31 7 2014	16 11 48
3	6	31 7 2014	16 14 55
8	20	31 7 2014	16 42 06
2	9	31 7 2014	16 50 00
9	35	31 7 2014	17 19 54
8	20	31 7 2014	17 27 56
1	15	31 7 2014	17 31 08
7	22	31 7 2014	17 37 10
9	35	31 7 2014	19 13 46
5	26	31 7 2014	19 40 12
8	7	31 7 2014	20 04 37
7	22	31 7 2014	20 06 21
9	13	31 7 2014	23 00 56
8	33	31 7 2014	23 07 41
9	32	31 7 2014	23 10 17

Figure 4.6: Example of a custom vocabulary and downloaded log from ADL Logger App. Each activity in the vocabulary has an ID and a description and is associated with a routine (1-9). The log has columns for the routine ID (R), activity ID (A), date stamp and time at which the activity started.

When an activity is logged, both the activity and routine ids are recorded along with the current date and time stamp. It is assumed that each activity starts when it is logged and ends when the next activity starts. This avoids the need for the user to log an end time or duration for each activity. Figure 4.6 shows an example of a log downloaded from the app. When the user selects 'download log' from the menu an email is automatically generated to be sent to the researcher. The email includes the current vocabulary and assigned unique IDs and each of the log entries since the log was last deleted. Each log entry is the routine ID (column R), activity ID (column A) and date-time stamp of when the start of the activity was logged.

4.2.3 Data collection and preprocessing

Using the custom built ADL Logger app, data was collected over sixteen non consecutive days for two healthy female volunteers (average age 27 years). The participants collected data on days when it was not too intrusive to affect their day to day lives between 16th July and 21st August 2014. Each participant created their own vocabulary during the collection process, which are listed in figure 4.6 and appendix A.1 for participant one and two respectively. It was found that it was easiest to start by entering a few common activities to the list and then add extra items to the vocabulary as necessary. The participants were also able to choose which of the pre-set categories they felt the activity best belonged to. Participant one had a final vocabulary of 40 activities and participant two had 37 activities, with 25 activities in common assigned to the same routine. Both participants also had 'sitting talking and 'waiting', but for participant one these were assigned to the 'relaxation' and 'travel' routines respectively, whereas participant two assigned both activities to 'multi'. Table 4.3 shows the number of activities each participant assigned to each routine. Comparing these vocabularies highlights the subjective and varied nature of participant defined labels.

Although it was not possible to edit the log on the app directly, if a participant made a mistake or realised they had forgotten to record an activity they made a separate note of the correct data and added this into the log once it had been downloaded. The logger only records the time of the start of a new activity and it is

Table 4.3: Number of activities assigned to each routine by participants

P	**No. activities per routine**								
no.	Wake up	Travel	Work	Lunch	Dinner	Exercise	Relax	Multi	Other
1	6	4	3	3	2	2	3	7	9
2	5	3	3	2	2	2	4	10	5

assumed that one activity continues until the next one starts. The number of seconds since the previous activity was calculated and vectors of the appropriate activity and routine labels were created at a 1Hz sample rate. This data was separated into time slices, equivalent of documents, using the same method and window length as for the UbiComp 08 dataset.

Initially, all of the days were processed together, however as each day is date and time stamped and the data was not collected on consecutive days this meant that the last activity of one day would be assumed to continue until the next time data was logged. This lead to these activities dominating the dataset as they would have a very large number of samples, particularly if there were a few days between the logs. To solve this problem it was necessary to process each day individually. However, this does mean that the last activity of the day, usually sleep has no data associated with it. This is similar to the UbiComp 08 dataset, where data was only collected during waking hours.

4.3 Detecting routines in the novel activity dataset

Topic models were estimated for the first 15 days of activity data for each participant. Trial and error of the number of routines and model parameters was used to establish the most semantically coherent model for each participant when evaluated qualitatively. Qualitative evaluation was performed by visualising the most probable activities in each discovered routine and judging the extent to which the combination of activities reflects a real routine. A leave-one-day-out cross validation was conducted using the selected number of routines and model parameters. For each day a topic model was estimated using the other fifteen days of data and then inference was performed for the left out day. The participants' grouping of their activities into routines when using the app were compared with the content of the routines found by the topic model to establish a mapping between them. The accuracy of the most likely predicted routine when compared to the mode of the ground truth labels for each 30 min time slice was calculated.

A quantitative assessment of the optimal number of routines was also conducted using a leave-one-day-out cross validation for different numbers of routines, from 5 to 50 at intervals of five. The perplexity (equation 2.15) of the left out day was calculated for every model and the average perplexity across all days was given for the range of routine numbers. The number of routines with the lowest perplexity is considered optimal [59], as a lower perplexity suggests a better generalisation performance.

The dataset collected contains several sources of bias, which should be taken into account when considering the routines detected. Firstly, the data was collected on non-consecutive days and the recording days were chosen arbitrarily by each participant in order to minimise the invasiveness of logging data. This results in an imbalance in the representation of weekday and weekend days in the dataset, which often include different routines. It also means that routines on days that participants considered it inconvenient to log data are not represented, which introduces an inherent bias as participants have specifically chosen not to include certain types of days. A bias was also introduced by removing long sleep activities due to not necessarily recording the waking time on the following day. Moreover, both participants were young, healthy females in full-time employment and thus only represent a small subset of the variability in routines and behaviour patterns across the full UK population. Finally, the participants had flexibility in the choice of activities logged and when 'unlabelled' was selected, which also adds an inherent bias in the activities represented.

4.3.1 Analysis and discussion of the detected routines

4.3.1.1 Routines discovered from models estimated for the first fifteen days

Using the first fifteen days of data, it was found that 12 routines gave semantically coherent results, which had good agreement with the ground truth labels for participant one. For participant two, 15 routines gave the most semantically coherent results. The participant's grouping of their activities into routines when using the app to log their data were compared with the content of the routines found by the topic model to establish a mapping between them, as shown in table 4.4. For participant two, even with 15 discovered routines, the activities that were logged under the 'exercise' routine are still mixed up with other activities and so no routine is discovered that maps to this ground truth label. This suggests that the labels forced by the app do not match the actual routine of the participant.

The routines found by the chosen topic models are shown at the bottom of figures 4.7 and 4.8 for participant one and two respectively. They are visualised as a list of the most probable activities for each of the routines, with the corresponding probability of the activity given in brackets. Inference was performed on the data for the sixteenth day using the topic models estimated for each participant from the first 15 days. The results of these were processed and visualised in the same way as for the UbiComp 08 dataset.

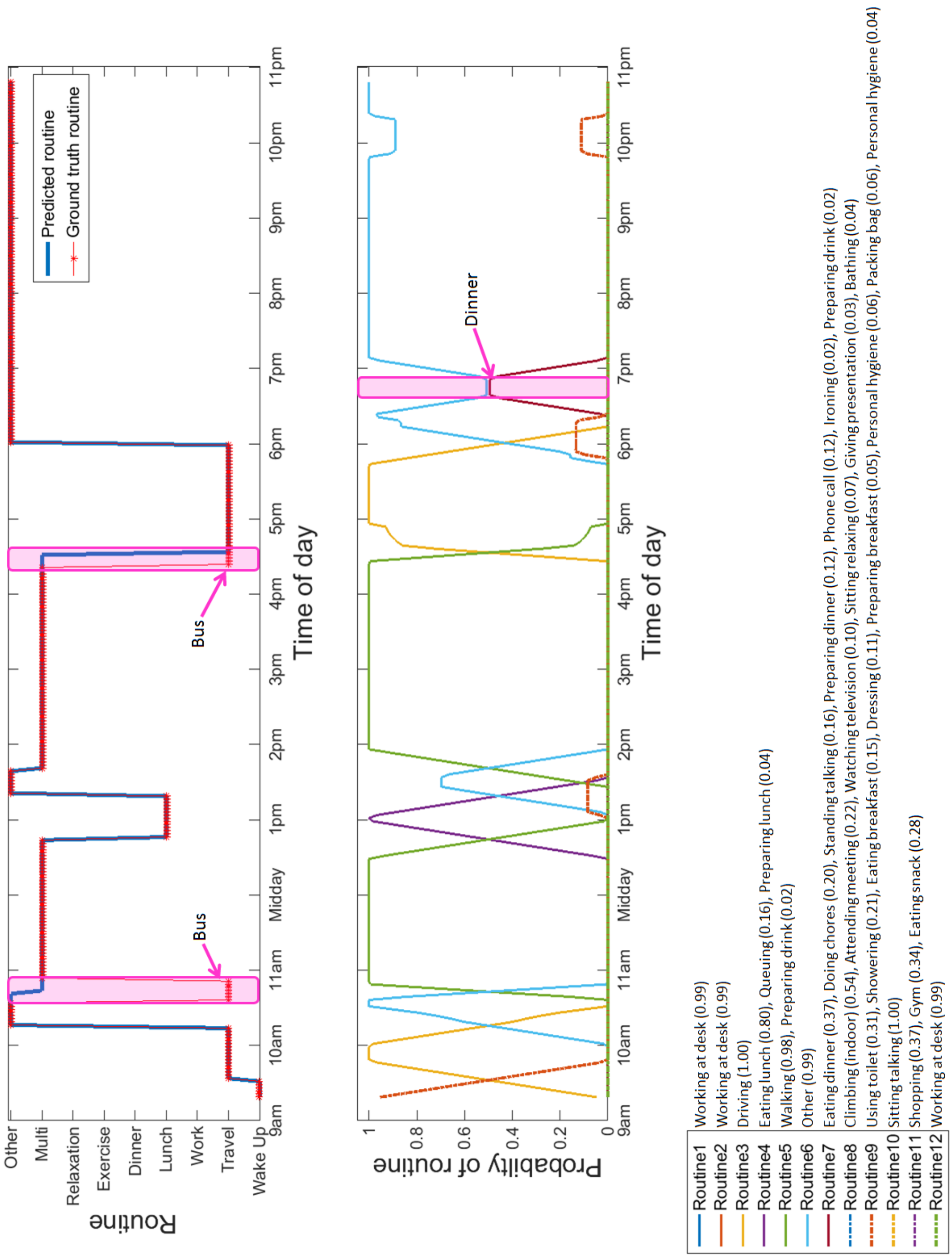


Figure 4.7: Results of Topic Model for Day 16 - Participant One

Table 4.4: Mapping of discovered routines from the models to ground truth routine labels from the logger app

Ground truth (App)		**Routine num (Model)**	
Label	ID	Participant 1	Participant 2
Wake up	1	9	14
Travel	2	3	4,9
Work	3	1, 2, 12	2,10,12,13,15
Lunch	4	4	5
Dinner	5	7	6
Exercise	6	8	-
Relaxation	7	10	3, 11
Multi	8	5	1,7
Other	9	6, 11	8

The top plot in figure 4.7 shows the actual logged routine and predicted routine for each document during the sixteenth day for participant one. It can be seen that overall there is a very good match between the ground truth recorded by the participant and the output of the topic model. However, there are some differences, in particular, the predicted routines have missed two bus journeys that occurred in the morning and afternoon that day, which were logged as part of the travel routine by the participant. This incorrect prediction is because there were no bus journeys in the previous 15 days of data that were used to create the model. Using a topic model with an infinite vocabulary may help to mitigate problems such as this where previously unseen activities occur [211].

Furthermore, it should be noted that the dinner routine has not been activated for either the predicted routines or the actual routines. Looking at the underlying activity data reveals that this was because on this day the only activity relating to dinner that was logged was ‘eating dinner’, which only lasted 5 minutes. Hence, this was not a sufficiently long enough event for it to be recognised because time slices are 30 minutes long and only the most common routine within that time period is recorded. This also happens for other activities that have a short duration, such as ‘using the toilet’. This is not necessarily a problem as the aim of using topic models is to investigate high level routines and the end users of the system could still look at the previous level of data fusion if they required more detailed information.

The second plot in figure 4.7 shows the probability of all routines for each time slice. This provides more information about the certainty of the predicted routines. In particular, this highlights that the ‘dinner’ routine is activated, but is not shown on the first graph as its probability is just below that of the ‘other’ routine. It can also be seen that routine 9, mapped to the ‘wake up’ routine occurs with a low

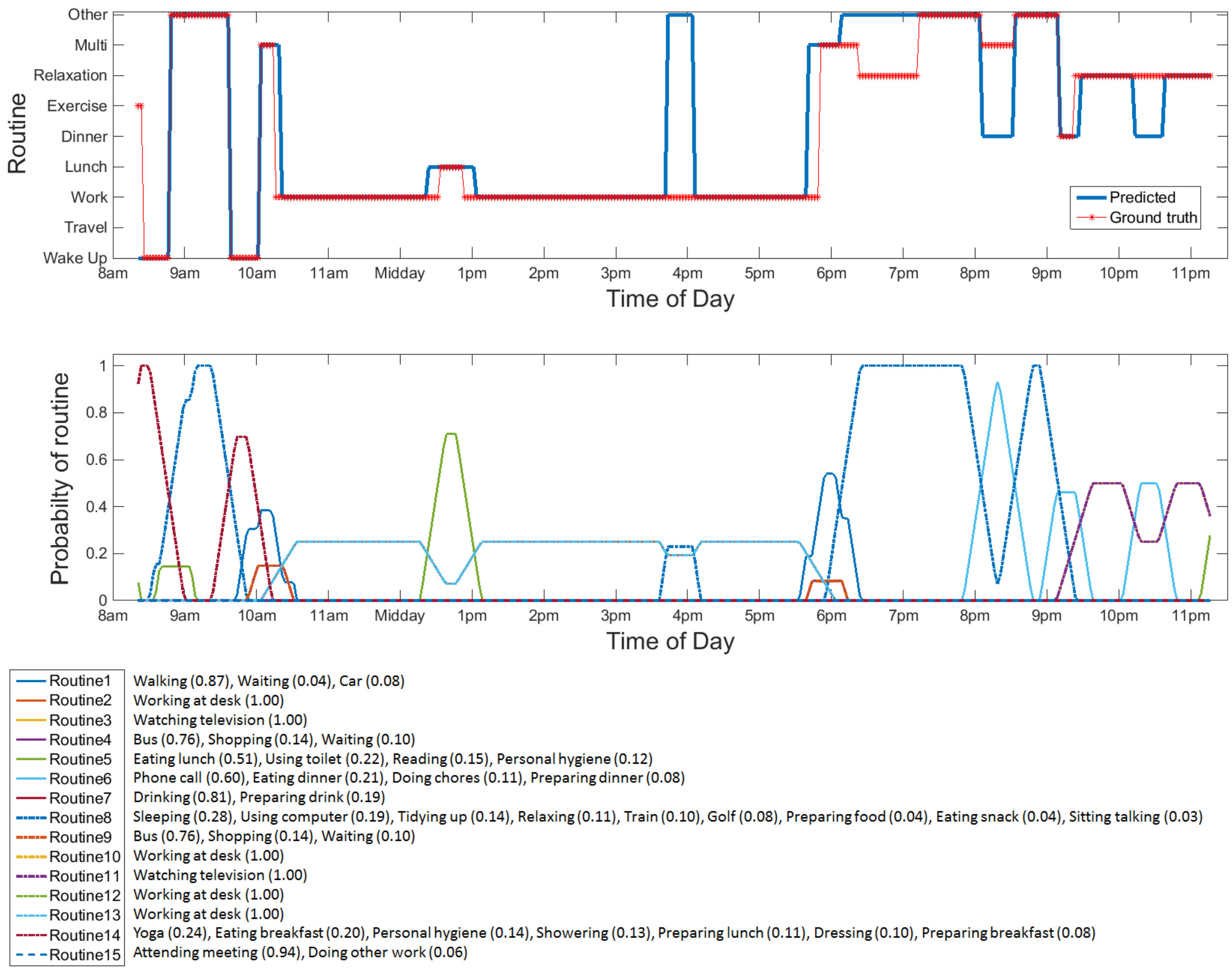


Figure 4.8: Results of Topic Model for Day 16 - Participant Two

probability at several points throughout the day. This is not necessarily intuitive but occurs because the model has associated the 'using the toilet' activity to this routine with a relatively high probability.

It can be seen from the top graph of figure 4.8 for participant two, that there is a good match between the actual and predicted routines for large periods of the day but that there are also several differences for shorter periods. The second graph confirms that often the routines are predicted with a relatively low probability, demonstrating uncertainty. One difference that can be seen is that the participant logged the 'exercise' routine first thing in the morning, whereas the topic model predicts the 'wake up' routine. However, this prediction is actually valid as the underlying activity is yoga, which is performed as part of the wake up routine but also classes as exercise. This highlights the limitations of the subjective nature of recording the ground truth and the limitations of the logging app. Similarly, some of the differences during the evening period could be considered valid. For example, routine 6, which has been mapped to 'dinner' includes the activities 'phone call' and 'doing chores', which is why it has been activated three times during the evening period. Hence, a more appropriate label for this routine may have been 'evening routine' but this was not an available option in the app and could be considered too generic to be useful.

In order to investigate the differences that occur between the actual logged routines and those predicted by the model, a matrix of routines was generated. Every time

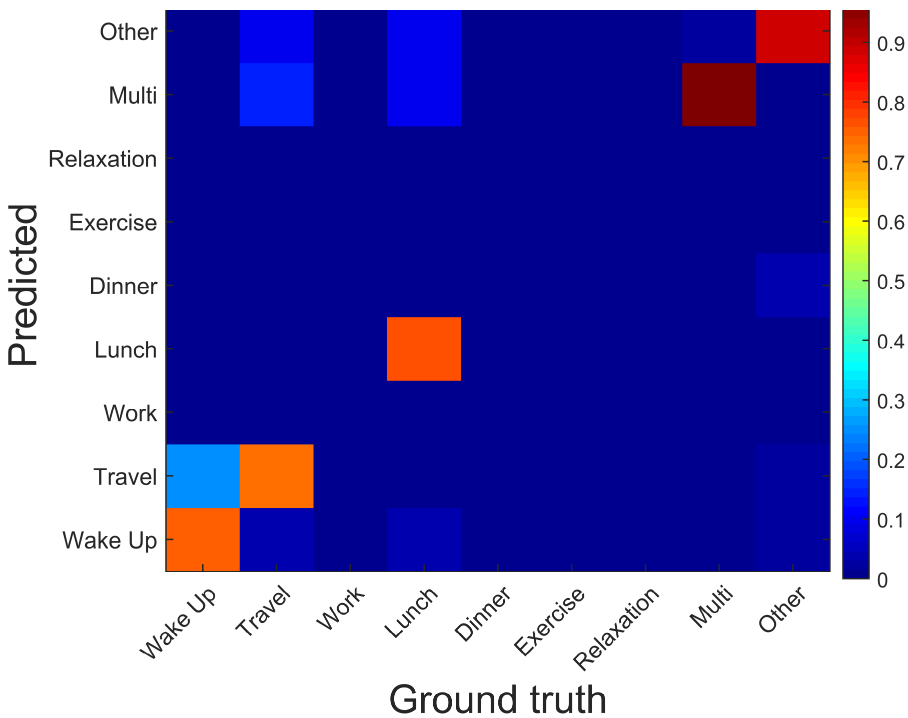


Figure 4.9: Visualisation of probabilities of predicted routines for each ground truth label (actual routine) for participant one.

slice was categorised by the mode of the logged routine labels during that period. For each routine label in the app the sum of the probabilities of all discovered routines (mapped to the relevant labels) for the corresponding time slices was calculated and normalised by the total number of time slices. Figure 4.9 shows a visualisation of this matrix for participant one, where the probability of the predicted routines for each ground truth label is displayed using a colour scale. The orange and reds across the diagonal show that, for the routines that occurred on this day, the probability that the ground truth label matched the predicted routine is high. The 'wake up' routine is mistaken for the 'travel' routine relatively often, this is likely to be because 'wake up' only lasts for a short duration, occurs once in the day and is immediately followed by 'travel'.

Figure 4.10 represents the matrix for participant two. This demonstrates that the probability of predicting a different topic is relatively high, particularly in comparison with the results for participant one. In addition, it can be seen that although exercise is logged it is never predicted as the topic model did not find a corresponding routine. Instead, the visualisation shows there is a high probability that exercise will be predicted as 'wake up', because it occurs in the morning.

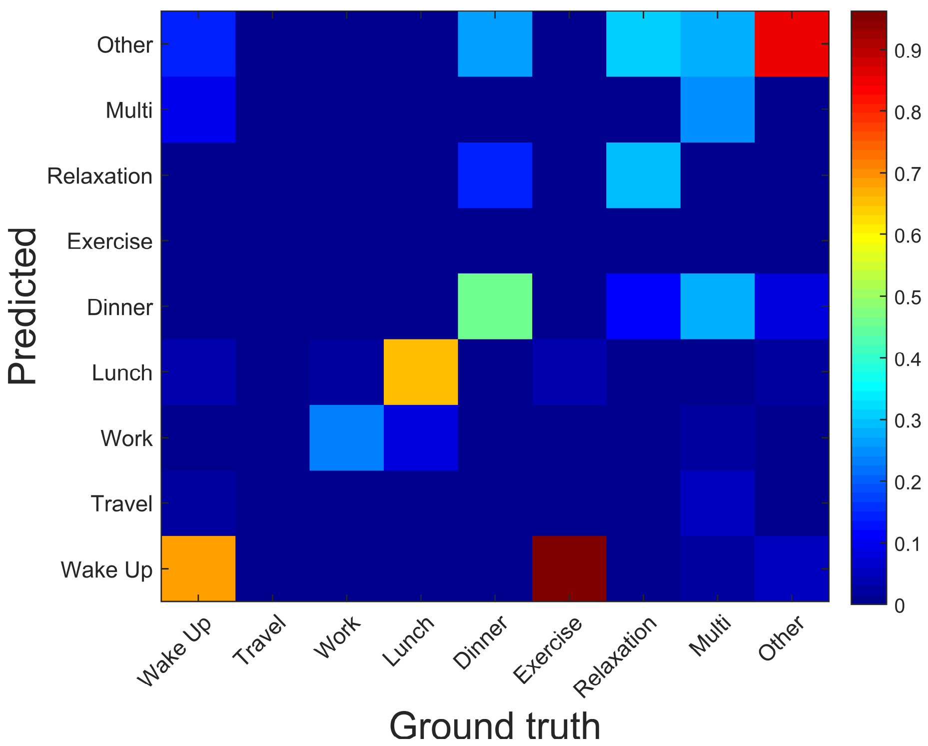


Figure 4.10: Visualisation of probabilities of predicted routines for each ground truth label (actual routine) for participant two.

4.3.1.2 Leave-one-day-out cross validation analysis

The results of the leave-one-day-out cross validation, using 12 routines for participant one and 15 routines for participant two, were used to calculate the accuracy of the models for each day. For each model the routines found were subjectively mapped to the ground truth routine labels in the same way as for the previous models. Using these mappings the most likely predicted routine for each time slice was calculated. These mapped predicted routines were directly compared to the mode of the ground truth labels for each time slice. The accuracy was calculated as the number of time slices with matching labels divided by the total number of time slices. The results for both participants are given in table 4.5.

Table 4.5: Accuracy of most likely predicted routine compared with ground truth labels for a leave-one-day-out cross validation for each participant

Day	**Accuracy P1** (%)	**Accuracy P2** (%)
1	33.3	89.7
2	91.9	76.3
3	83.1	70.4
4	90.1	78.5
5	64.0	64.5
6	88.2	41.8
7	91.4	81.6
8	87.2	80.3
9	84.8	76.5
10	79.3	80.9
11	91.5	53.8
12	77.5	38.2
13	84.3	87.4
14	78.3	70.3
15	88.0	47.6
16	62.8	67.4
Average	79.7	69.1

The results confirm that overall there is a greater level of uncertainty in the models for participant two, with an average accuracy of 69.1%, than there is for participant one, with an average accuracy of 79.7%. This is probably because the ground truth labels matched the behaviours of participant one more closely. This highlights that the subjective nature of the ground truth labelling can affect the results, even though the routines found may be realistic. For participant one, day 1 has a much lower accuracy than any other day. This is because the participant only recorded the activity of ‘gym’ on this day. Therefore, as the model was trained on the other days when this activity did not occur it was not possible for the correct routine to be predicted. Moreover, day 1 for this participant only started in the afternoon and therefore the 'gym' activity accounted for quite a large percentage of the data for the day, making the impact of this problem larger.

For participant two, day 6 and 12 have the lowest accuracies. This is probably because both of these days were Sundays, whereas all of the other days in the dataset were weekdays. The participant's activities and routines are very different at the weekend and there is little training data, therefore the model does not perform as well for these days. Investigating the discrepancies in the predictions and the ground truth confirms this. For example, on day 6 the participant takes a long train journey, however the corresponding activity only occurs in routine 6 with a probability of 0.06 and therefore this routine has been mapped to 'Other' rather than 'Travel' as it contains several activities with much higher probability that are related to the 'Other' routine.

4.3.1.3 Quantitative analysis for selection of number of routines

The results of the leave-one-day-out cross validation across different numbers of routines were used to calculate the average perplexity across the range for each participant. Due to the problem of the 'gym' activity not occurring in any of the training data when leaving out day 1 for participant one the perplexity for this day is of the order of magnitude 10^9 or 10^{10}. Therefore this day has been excluded from the

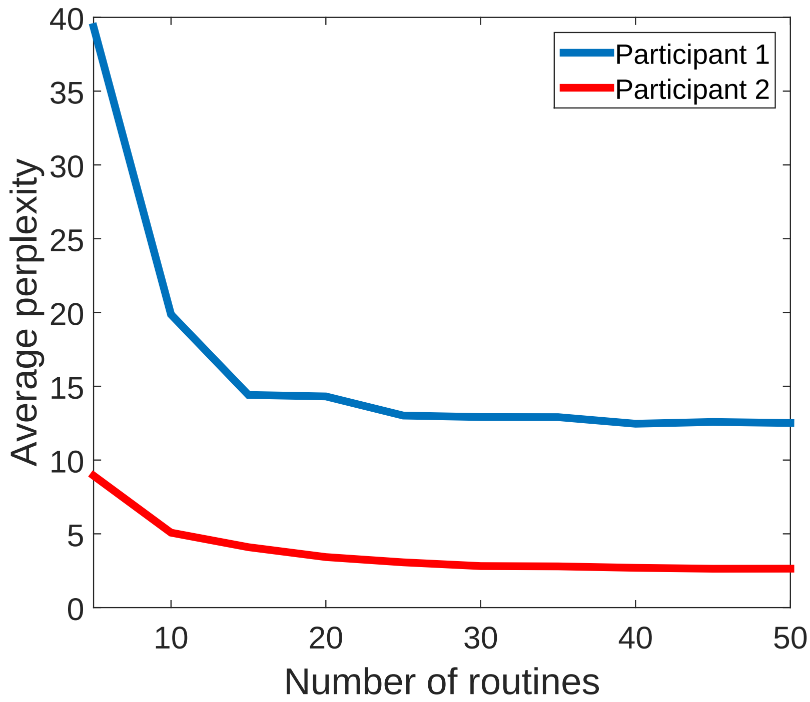


Figure 4.11: Average perplexity of models for different numbers of routines

average for participant one as it obscures the remaining results. Figure 4.11 shows how the average perplexity varies with the number of routines. It can be seen that overall the perplexity is lower for participant two, which suggests that the models for this participant have a better generalisation performance. This is contrary to the accuracy results, highlighting that qualitative and quantitative assessments do not always correspond due to the subjective nature of the labelling process. Labelling the discovered routines for each participant based on their content could provide more realistic results than using a standardised set of ground truth labels.

The average perplexity for participant one has a high rate of change for less than 15 routines. For participant two the rate of change is steadier, reaching approximately zero at 20 routines. This suggests that models with slightly higher numbers of routines would perform better than the 12 and 15 routines, that were subjectively chosen to give the most semantically coherent results. However, although a model with 20 routines may generalise better, it would not be as useful for further analysis of daily routines as the size of the vocabularies for these models is only approximately double this number of routines. Therefore the routines would be too specific and be strongly associated with single activities rather than groups that form a routine. Therefore, the semantic coherence is important when selecting the number of routines to use.

4.4 Limitations of data collection method

When using the initial version of the ADL Logger app some shortcomings were identified. In particular, it was noticed that some activities are carried out concurrently, for example the user may be working at their desk whilst drinking a coffee or watching television whilst talking on the phone. If the user has chosen to enter the activities they can select from individually, i.e. ‘drinking coffee’, ‘working at desk’, ‘watching television’ and ‘talking on phone’ then they will need to make a subjective decision as to which the main activity currently is and log that one. This can mean that data that is of medical interest, such as fluid intake or social interaction, is lost. One solution is to list all combinations of activities as separate items in the vocabulary. However, this might result in a large vocabulary, making logging activities more time consuming and requiring a larger dataset to ensure sufficient examples of each activity are included. On the other hand, it is expected that some detail will be lost when summarising data at a higher level and this is not necessarily a problem. If clinicians or users require more detailed information then the system could allow them to access the results from lower levels of data fusion.

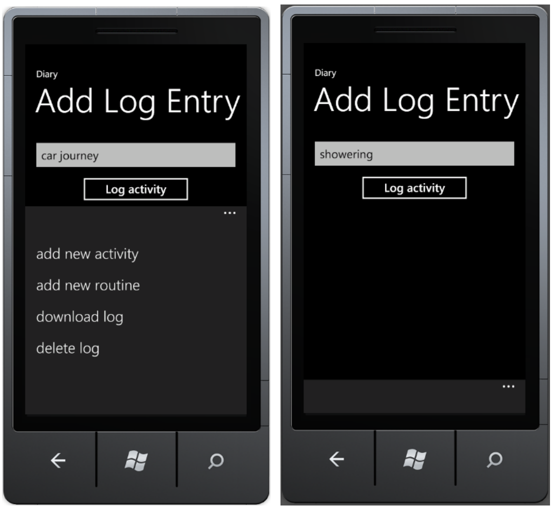


Figure 4.12: ADL Logger App (a) Menu (b) Adding a log entry

Another limitation of the app is the lack of functionality to correct mistakes, e.g. when a user forgets to record an activity or accidentally logs an incorrect activity. This problem was addressed by post-editing of the log from the user's memory of events to provide as realistic a representation as possible. Furthermore, when adding activities to the vocabulary for the ADL Logger app the routine options to select from are restrictive, even with the 'multi' option and do not always reflect the user's behaviour patterns. For example, participant two added the activity 'yoga' to the 'exercise' routine, however this activity actually generally occurs as part of the morning routine and hence it may be more appropriate for it to be in the 'wake up' routine. This subjective selection of routines can mean the results of the topic model do not always agree with the ground truth, even if they are valid. Finally, the feedback to indicate an activity has successfully been logged after pressing the '+' button was too subtle.

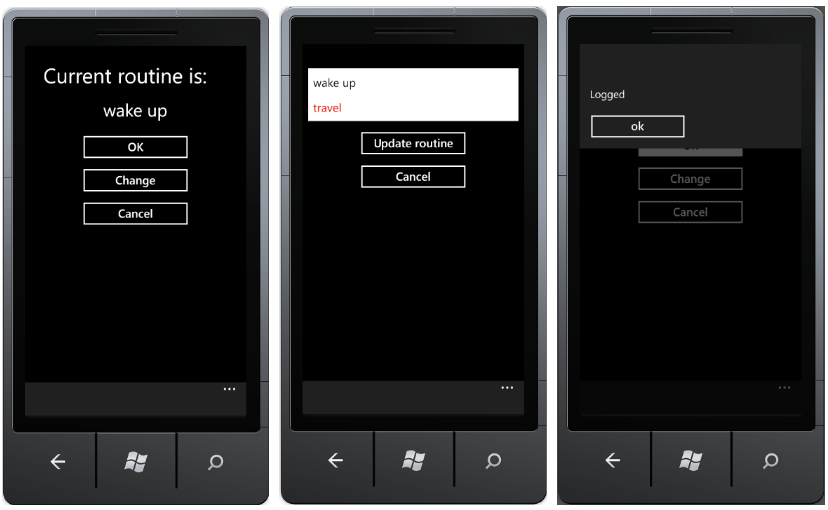


Figure 4.13: ADL Logger App (a) Check topic (b) Update topic (c) Confirmation

The ADL Logger app was improved to address the restrictive choice of ground truth routines. Instead of having a pre-defined set of routine labels, the user can use the menu options, shown in figure 4.12(a) to separately add descriptions of different activities and routines. A new activity can then be logged by selecting it from the drop down list and pressing the ‘Log activity’ button, as shown in figure 4.12(b). Based on the existing assumption that a routine runs over a period of several activities, when the user logs an activity they are asked whether it is still part of the current routine or if it is the start of a new routine, as shown in figure 4.13 (a). If the new activity is the start of a new routine then the user presses the change button and can select the new routine from the list, as shown in figure 4.13 (b) and then press update routine. Once an activity has successfully been logged a confirmation message is shown to the user, as seen in figure 4.13 (c), to provide clear feedback.

4.5 Chapter summary

This chapter considered the application of LDA to activity data with the aim of discovering daily routines. The work of Huynh et al. [32] was replicated and the results for the UbiComp 08 dataset were shown to be similar to those previously published. In particular, the topic model with 10 routines had an average accuracy of 81.9% when the most likely predicted routine for each time slice was compared with the corresponding ground truth label for the results of a leave-one-day-out cross validation. The structure of the routines discovered by the topic model were comparable to those found by Huynh et al. in terms of the most probable activities.

The requirements and challenges for collecting an ideal dataset for daily routine detection from activity data collected in a residential environment were discussed. The development of a mobile phone application to collect a new dataset of activities and ground truth routine labels was described. The experience of collecting data and the problems encountered, such as concurrent activities were highlighted and discussed, with suggestions for improvements given. LDA was applied to the activities in the novel dataset collected to discover patterns reflective of daily routines. The results presented in this chapter have shown that topic models can be successfully used with different activity datasets to discover daily routines. On average, across both datasets, there is a high level of accuracy, approximately 70 to 80%, when comparing the discovered routines with ground truth labels. Moreover, each of the datasets included different activities, showing that topic models will work with varying vocabularies.

For the new dataset, the results varied between the participants, in particular there was a higher level of uncertainty in participant two but a lower average perplexity. This highlights that the subjective nature of both ground truth labelling and assigning labels or mappings to the discovered routines can have a large impact on the results and need to be considered carefully. The semantic coherence of the routines may be of more importance than how well they correspond to the ground truth labels. This is particularly true due to the unsupervised nature of topic models. For example, when discovering a routine that the user has not considered labelling but the content relates to a routine the user performs the results of the topic model could be considered more accurate than the original ground truth.

Finally, the results have also highlighted that the amount and variation of the data used to estimate the model can have a large impact on the inference results on a held-out dataset. For example, having an activity occur for the first time in a held-out dataset will mean that the model is unable to accurately determine the corresponding routine, especially if the activity lasts for a long time. One solution to this could be to use a topic model with an infinite vocabulary [211]. In addition, a very different routine structure in a held-out day, for example, a weekend day vs a weekday will reduce the inference performance if the model was estimated on data without similar routine structures.

Chapter 5

Changes in Routines Over Time

Chapter 4 demonstrated that Latent Dirichlet Allocation (LDA) can be successfully applied to activity data to discover daily routines for an individual. The routines found are based on the data used to estimate the model and hence are the most probable combinations of activities for that period of time. However, in reality many peoples' routines are not fixed and will change over time, this is an example of concept drift [212]. These changes are two-fold: the amount of time spent in each routine fluctuates over time and the probability of the activities that occur within a routine may also vary. For example, an individual may change their job and have a shorter commute so the amount of time spent doing this routine would decrease but the activities involved in the routine would remain the same. Alternatively, an individual may change their mode of commuting from driving to cycling and hence the probability of the activities in the 'commuting' routine will alter.

The two commuting examples given are dramatic shifts in behaviour patterns in order to highlight the two types of changes. However, more subtle changes over time, that would not necessarily be noticed by an individual or their clinician, may be an indicator of an alteration in health status. Being able to detect such variations could help identify concerns and aid earlier diagnosis or monitor improvements in response to treatment. For example, someone with the onset of depression may start to spend less time preparing meals as they have less motivation and energy, hence changing their habits from cooking home-made meals to eating ready meals. This would mean the probability of the activities included in their dinner routine would alter over time.

Dynamic topic models (DTMs), introduced in section 2.3.1, can be used to identify the second type of change, that is when the probability of activities in a routine vary over time. This chapter investigates how DTMs can be applied to long term daily activity datasets to detect these changes. There are no datasets publicly available with annotations of changes over time, therefore simulated datasets are used as this allows the changes to be controlled. Initial experiments were conducted to investigate different approaches for applying DTMs to activity data, to detect changes over time in the probabilities of activities within a routine. A validation experiment was then conducted to establish whether specific changes can be correctly identified without prior knowledge of what the changes are.

5.1 Applying dynamic topic models to long term activity data

Dynamic topic models are an extension of LDA that allow for the evolution of routines over time. Unlike LDA where documents are assumed to be exchangeable within the entire corpus, DTMs impose a sequential structure on the data. This means that changes in the probability of the activities in each routine over time are explicitly captured using this model. This is achieved by splitting the dataset into ordered time slices, where a noticeable change could be expected to occur between one time slice and the next, for example a month. The data in each large time slice is used to create smaller time slices, for example 30 minutes of data, as done in section 4.1.1 for the LDA model. These 30 minute slices of data are the equivalent of documents in the original text processing context. For each large time slice an LDA model is estimated but the routines found are constrained to evolve from one time slice to the next, as shown in figure 5.1. More specifically, the routines, β_k, are chained between the large time slices using a state space model that evolves with Gaussian noise.

5.1.1 Dynamic topic model implementation

The approximate variational inference method for dynamic topic models from Blei and Lafferty [53], summarised in section 2.3.1, was implemented in C++ by Sean Gerrish and David Blei along with the Document Influence Model [213] and released as open source code [214]. An executable binary of this program is available on Github [215] and was downloaded for use. A python wrapper for the C++ program is included as part of the open source library Gensim, created by Radim Řehůřek [216]. This python library provides functions to create and edit a corpus, call the DTM program and view the resulting topics for each large time slice. The settings and parameters for the DTM are passed as arguments to the C++ program. The python library has default values for many of these settings, which were used for standard parameters, such as maximum number of iterations of the EM algorithm. For each

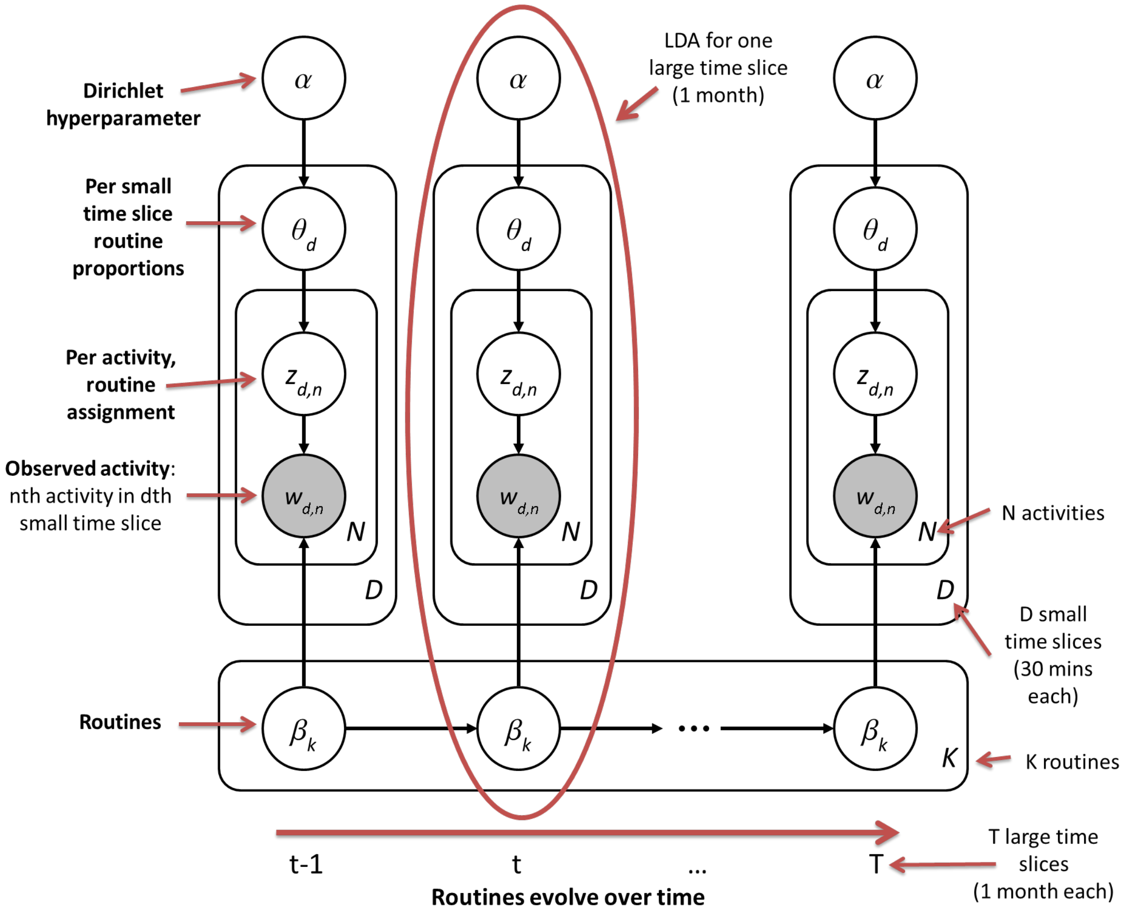


Figure 5.1: Applying the dynamic topic model to long term activity data

model the dataset used, length and number of large time slices required, number of topics, vocabulary and alpha values were specified.

Preprocessing of the dataset to create a corpus in the same format as described in section 4.1.1 for LDA was conducted in Matlab, as this had been used for the previous work, thus allowing code reuse where appropriate. The Gensim python library can be used for simple visualisations of the top activities in routines over time. However, more detailed analysis of the results was conducted in Matlab by directly using the data files output by the C++ program as this provided greater flexibility than interfacing through a python API.

5.1.2 Simulation of long term activity data

To date there are no open access long-term datasets for activities of daily living that are annotated with changes in routines over time. Therefore, it is not possible to directly validate changes in the activities of a routine discovered using dynamic topic models on a real world dataset because it is not known if these variations actually occurred. Although simulations of activity datasets are oversimplified and not always realistic, they offer the benefit of controlling specific parameters of the dataset. Therefore simulations can be used to generate a long-term dataset which has a known change occurring in it. The changes in routines discovered by a DTM can be compared to the known change that was included in the simulated dataset to investigate whether it is possible to use this model to detect changes in routines over time.

There are a variety of smart-home simulators in the literature [217–220], however at the time of writing many of these do not have code available to use for running simulations. Moreover, several of these simulators are designed for different applications, such as designing smart homes and hence are not suitable for the purpose of creating a long-term activity dataset. Often simulators require a high level of user input, such as paths taken around the environment by the simulated humans or a list of pre-defined events. The output of the simulators also varies, with several implementations only giving logs of simulated sensor activations. For this experiment, output at the activity level is desired in order that this data can be directly used as the input for the dynamic topic model to discover routines in the activities and how they vary over time.

Given the requirements, the chosen simulator was the Home Sensor Simulator created by Kormanyos and Pataki from Budapest University of Technology and Economics [221]. The software, written in C#, is freely available for use and includes two small programs that allow different layouts of houses and human behaviour profiles to be created. These can be used with the main simulator to generate data as required. A variety of different data logs are generated for each simulation, including activity logs at two different levels of abstraction. The output from the simulation also includes detailed sensor activation logs for all simulated sensors placed in the house layout and movement data for the simulated human. The simulation length and step size can be specified and the results are displayed as graphs and saved to file.

Kormanyos and Pataki [222] created the Home Sensor Simulator to help develop and test algorithms for ambient assisted living projects without the need to collect real life data with ground truth labels. The model assumes that there is a single human inhabitant, with specified behaviour characteristics, in the simulated home environment. The actions and activities that occur are decided based on priorities, which are affected by the state of the environment and the behaviour profile. For example, someone will become thirsty more quickly if the temperature is high, they are performing a physical activity and have not had a drink recently. The simulator assumes that exactly one activity, at the highest level of abstraction, is occurring at any given time and each of these activities has an associated priority function. The activity with the highest priority will take precedence. Priority functions are based on the combinations of relevant constants and variables describing the current situation, including an element of randomness that reflects the complex nature of real life behaviours.

The hierarchical approach to defining activities allows flexibility in the level of detail for each activity. High level activities are made up of a sequence of middle level activities. For example, the *eat_cold_meal* event includes *get_ingredients, eat_meal_at_table* and *put_plate_to_sink* [222]. Unlike the high level activities, those in the middle level do not have a priority function. The middle level activities are further subdivided into actions, which directly influence the variables associated with the simulation and hence affect the priorities. For example, *eat_meal_at_table* is split in to the actions *go_to_table* and *eat_meal.* Furthermore, most actions can be interrupted by others and then resumed once the interrupting action has been completed. Interruptions are controlled by the current interrupt level, based on a combination of the priority level of the main activity and the interrupt levels of the corresponding lower level activities and actions currently being performed. The naming convention for activities is taken directly from the simulator and are generally self explanatory. At the low level, moving from one labelled area in the house layout to another is prefixed by the word 'go'. A full list of activities at the low and high levels of abstraction are given in appendices A.2 and A.3 respectively.

For all of the experiments described in this chapter, the Home Sensor Simulator was used to create long term daily activity datasets using the default large home layout. The parameters for the human behaviour profiles are given as a percentage and the real world meaning of the range of variability is detailed in table 5.1. A simulation is run to generate one month's worth of data and then the human behaviour parameters are adjusted, as detailed in each relevant section. This is repeated until data representing a full year is simulated, with a known gradual change occurring. For each dataset, a total of 12 months were simulated and all months were set to be 30 days, with a simulation step size of 10 seconds.

5.2 Investigating the effect of activity data properties on detecting changes

This section describes initial experiments to investigate the performance of using a dynamic topic model to find a known change in a simulated dataset. The aim is to find the change occurring at the activity level, that is, the probability of the

Table 5.1: Human behaviour profile parameters available for the Home Sensor Simulator. Every parameter is given as a percentage and the real world corresponding ranges are in italics. The chosen values for simulations 1 and 2 are listed.

#	Parameter	Simulation 1 / %	Simulation 2 / %
1	Eat frequency *3 to 7 meals per day*	25 to 80 (+5/month)	50
2	Eat warm frequency *0 to 1 meal per day*	50	50
3	Drink frequency *rarely to often*	50	50
4	Sleep lengths *5 to 10 hours*	50	50
5	WC use frequency *rarely to often*	50	25 to 80 (+5/month)
6	Move speed *0 to 1 m/sec*	50	50
7	Shopping frequency *never to often*	50	50
8	Visitor frequency *weekly to 3 times a day*	50	50
9	TV frequency *never to often*	50	50
10	Computer frequency *never to often*	50	50
11	Computer lengths *5 mins to 2 hours*	50	50
12	Exercise frequency *never to often*	50	50
13	Go outside frequency *never to often*	50	50
14	Go outside lengths *20 mins to 2 hours*	50	50
15	Wash dishes *never to often*	50	50

relevant activities should change over time within one or more routines. Changes can also occur at the routine level, where the probability of the whole routine occurring changes but the activities within the routine are not affected, these types of changes are not the focus of this work. Two simulations were run, each changing one parameter each month, whilst keeping all other variables constant. The data generated by these simulations were processed to create long term activity datasets with different properties that were used as the input to dynamic topic models. The results were explored through visualisations and analysed to determine if the known changes could be detected.

5.2.1 Experimental methods for investigating the effect of activity data properties

5.2.1.1 Modelling known changes over time

For the first simulation, the eating frequency behaviour parameter was changed by 5% each month, whilst all the other parameters were kept at 50% as this reflects average behaviour in the simulator, as detailed in table 5.1. The hypothesis was that activities associated with eating would increase in probability within one or more routines. Using the method described in section 5.1.2 a year of simulation data was generated. A sample of a full event log generated by the simulation software, containing activities at the lowest level of abstraction, is shown in table 5.2.

Table 5.2: Full event log sample with activities at a low level of abstraction

Day No.	**Time**	**Activity**
1	00:00:00	go_wardrobe
1	00:00:24	get_clothes
1	00:01:33	go_bathtub
1	00:01:51	have_bath
1	00:26:22	go_bathroom_sink
1	00:26:28	brush_teeth
1	00:31:56	go_bed

The vocabulary for the dynamic topic model was established to be all unique activities that occurred in the full event logs for the 12 month simulation. The event logs were processed to prepare the data as input to estimate a dynamic topic model. It was assumed that the current activity continues until the next activity in the log starts and that activities are cut off at the end of the month. Activities were listed for each second (1Hz) for the full 12 months. For consistency with the method used for LDA, in section 4.1.1, 30 minute time slices were created from this 1Hz activity data with a sliding window of 2.5 minutes. Each month was considered a large time slice for the DTM, giving 17269 small 30 minute time slices per large time slice. Dynamic topic models were estimated using the data generated with 12 large time slices, a vocabulary of 76 activities and $\alpha = 0.1$.

The top words and their probabilities across all of the months, for each of the routines, were viewed using Gensim [216]. Every routine was assigned a label based on the top activities associated with it. The discovered routines were visualised and explored in different ways to help identify significant patterns, as summarised in section 5.2.2.1, to see if they agreed with the hypothesis that eating related activities would become more likely in one or more routines. The raw simulation results were also analysed to ensure the induced change had been simulated as expected. Finally, the most likely routine for each 30 minute time slice was determined from the results of the DTM. A histogram of the routines for each month was generated and the values for each routine across all months were normalised to be in the range 0 to 1.

For the second simulation, the same methods were used but instead of changing the eating frequency behaviour parameter, the WC use frequency parameter was changed by 5% each month, as detailed in table 5.1. Similarly to the first simulation, the hypothesis was that activities associated with using the toilet would increase in probability within one or more routines. Generally, using the toilet is an activity which can occur as part of many routines at any time of the day and has a relatively short duration. Therefore, it was expected that these changes would be more obvious at the activity level than those related to eating.

5.2.1.2 Impact of activity durations

The exploratory visualisations and analysis of the results for the two simulations, detailed in section 5.2.2, revealed that detecting changes in the probability of activities within a routine over time is not trivial. Changes in activities of longer durations, such as eating, occur at the routine level rather than within a routine. This problem could be addressed by increasing the length of the small time slices, which is currently 30 minutes. However, this would mean that short activities would comprise a smaller percentage of each small time slice and hence not be well represented by the model.

To understand the length of the different activities, the mean, maximum and minimum duration of each activity at the low level of abstraction in the second simulation was calculated, as shown in figure 5.2. This reveals that the mean duration of all but two of the activities is less than one hour. The ‘walk outside’ activity has a mean duration of 01:12:31, which is only slightly greater than an hour. In contrast, the ‘sleep in bed’ activity has a mean duration of 07:33:24, which is significantly longer than any other activity. The high mean duration for this activity causes a large number of the small time slices used as the input to the dynamic topic model to only contain this one activity. Therefore the changes in other routines that are of interest are only represented in a relatively small proportion of the small time slices. This under-representation reduces the ability of the dynamic topic model to identify the changes of interest within routines.

To determine the impact of the long sleep activity on the performance of the dynamic topic model it was removed from the simulated dataset for the second simulation, where the frequency of the WC use parameter was increased with time, as

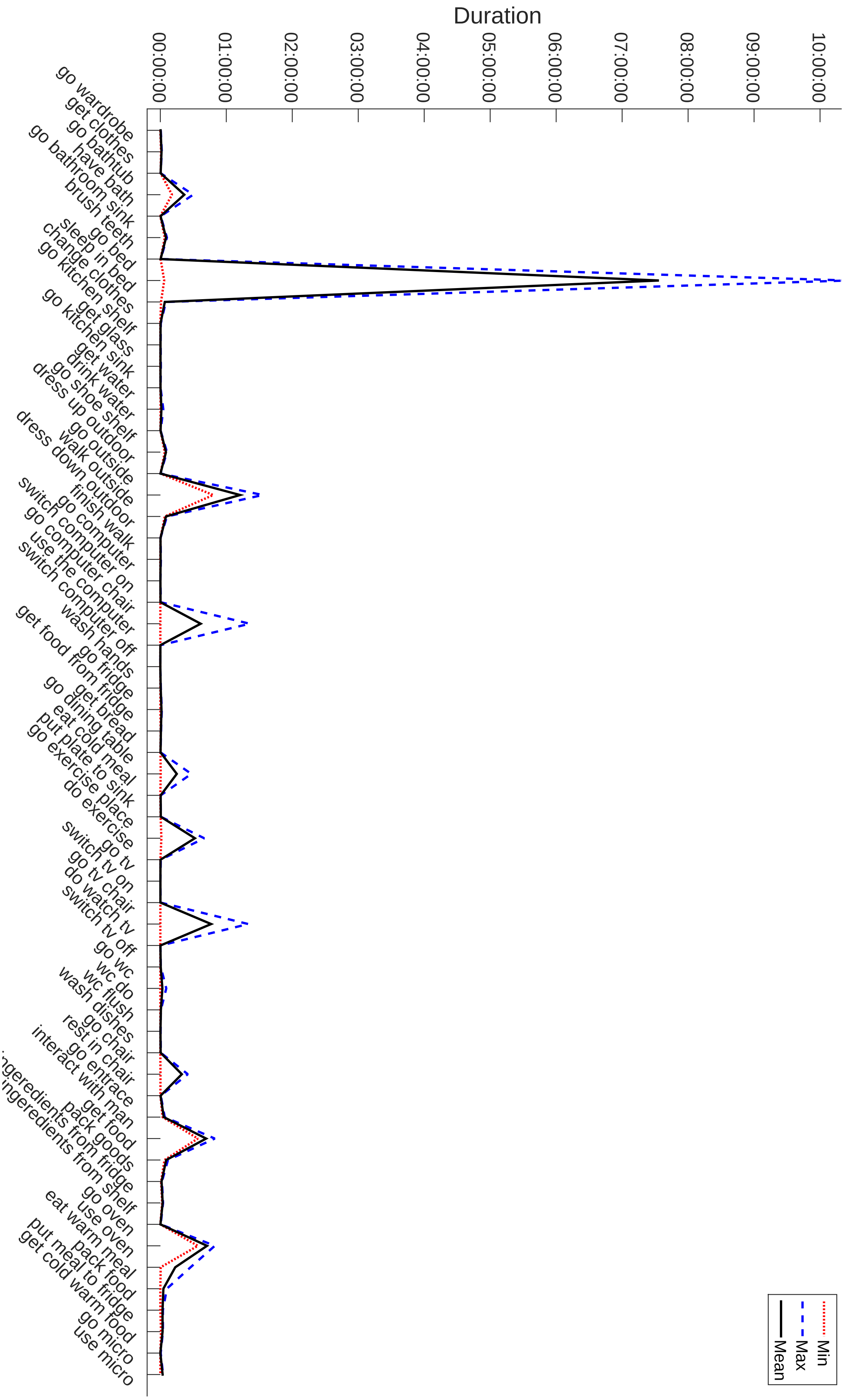


Figure 5.2: Minimum, maximum and mean durations of all activities at the low level of abstraction for simulation 2.

this change was found to occur at the activity level. As the mean duration of the majority of activities is less than an hour, any occurrences of the 'sleep_in_bed' activity that lasted longer than 1 hour were removed. This meant that short naps would still be captured as these are different to sleeping overnight. As it is assumed that each activity in the log continues until the next one starts the dataset was split in to days, with the last activity ending when the first long sleep activity began. Short bursts of activity, such as getting a glass of water, in between long sleep activities were discarded if they were shorter than the window length.

Another factor related to the duration of activities is the choice of vocabulary, i.e. the level of abstraction of the activities used as input data. Therefore the duration of the activities at the highest level of abstraction, listed in the shorter event log were also investigated. The mean, maximum and minimum duration of each high level activity in the second simulation was calculated, as shown in figure 5.3. There is only one activity (do the dishes) with a mean duration less than 1 minute, whereas 56% of the low level activities are under 1 minute. Similarly to the low level activities, the sleep activity has a much longer mean duration, 08:04:17, than any other activity.

Using the high level activities with longer durations means each activity comprises a larger proportion of each small time slice. Therefore changes in the probability of activities within a routine will be more obvious than a small change occurring for every short activity related to watching the TV, for example. However, the disadvantage

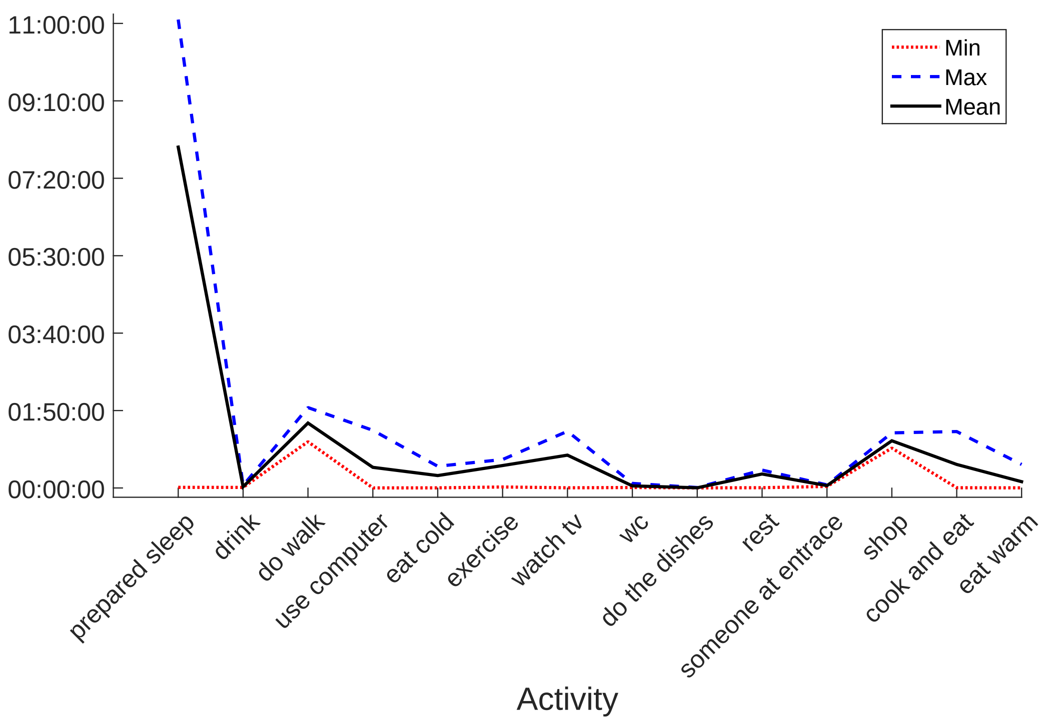


Figure 5.3: Minimum, maximum and mean durations of all activities at the high level of abstraction for simulation 2.

of using a smaller vocabulary of activities at a higher level of abstraction is that they become more like routines. Hence some routines can be comprised of a single activity with probability of 1.0, providing no further insight than the original data. To determine the impact of using activities at a higher level of abstraction, the shorter event logs were used to generate the 30 minute time slices as input to estimate a dynamic topic model with 12 large time slices, a vocabulary of 14 activities and $\alpha = 0.1$. Finally, combining the use of the high level activities with the removal of long sleep durations was investigated.

A total of five datasets were generated from the two simulations to investigate the effect of the activity data properties, as summarised in table 5.3. Only one dataset was based on the simulation where the eating frequency parameter was changed because it was found that this change occurred at the routine level and this is not the focus of this work. Four datasets were based on the simulation where the WC use frequency parameter was changed to investigate the effect of different combinations of properties related to the duration of activities.

Table 5.3: Summary of properties for datasets generated from simulations with simulation parameter changed over time, whether long duration sleep activities are included and activity abstraction level.

ID	Simulation parameter	Long activities	Abstraction level
A	Eating frequency	Yes	Low
B	WC Use frequency	Yes	Low
C	WC Use frequency	No	Low
D	WC Use frequency	Yes	High
E	WC Use frequency	No	High

5.2.2 Visualising changes in routines

This section highlights key results of the exploratory investigation to determine the effect of the activity data properties on the performance of a DTM. A variety of different visualisation techniques are considered to facilitate interpretation of the results.

5.2.2.1 Results of changing eating frequency parameter

The results in this section relate to dataset A where the eating frequency parameter was changed. The top three activities and the corresponding range of values for their probabilities over 12 months for each routine discovered by a 10 topic DTM are shown in table 5.4. Each routine is arbitrarily numbered 1 to 10 and assigned a subjective label to describe the routine based on its constituent activities. The results show that the probabilities of the top activities in most routines are static or have a very low variation, less than 0.005.

Although there is little variation in the probabilities of activities in routine 5, the order of the top activities does alter over time as one activity becomes more probable than another. This concept of changes in the importance of activities in

Table 5.4: Top activities and their range of probabilities over time for routines discovered using dataset A, changing eating frequency. Each routine is given a subjective label. A dynamic topic model with 10 routines over 12 months was used.

Routine No.	**Routine Label**	**Top 3 Activities**	
		Description	Probability range
1	Sleeping	sleep_in_bed	1.000
2	Cooking	use_oven	0.921 - 0.922
		get_ingredients_from_shelf	0.046
		get_ingredients_from_fridge	0.024 - 0.025
3	Exercise and hygiene	do_exercise	0.548 - 0.564
		have_bath	0.305 - 0.317
		brush_teeth	0.073 - 0.077
4	Eating and social	eat_warm_meal	0.405 - 0.510
		interact_with_man	0.192 - 0.221
		pack_food	0.071 - 0.082
5	Relaxing	rest_in_chair	0.901 - 0.904
		wc_do	0.018 - 0.020
		drink_water	0.018 - 0.019
6	Outside	walk_outside	0.996 - 0.998
		dress_down_outdoor	0.001 - 0.002
		dress_up_outdoor	0.000 - 0.001
7	Getting ready	change_clothes	0.299 - 0.325
		dress_up_outdoor	0.151 - 0.166
		dress_down_outdoor	0.149 - 0.165
8	Relaxing / TV	do_watch_tv	0.972 - 0.973
		drink_water	0.011
		go_tv	0.005
9	Using computer	use_the_computer	0.960 - 0.961
		drink_water	0.015
		go_kitchen_shelf	0.006
10	Eating	eat_cold_meal	0.593 - 0.649
		get_food	0.130 - 0.148
		get_food_from_fridge	0.036 - 0.048

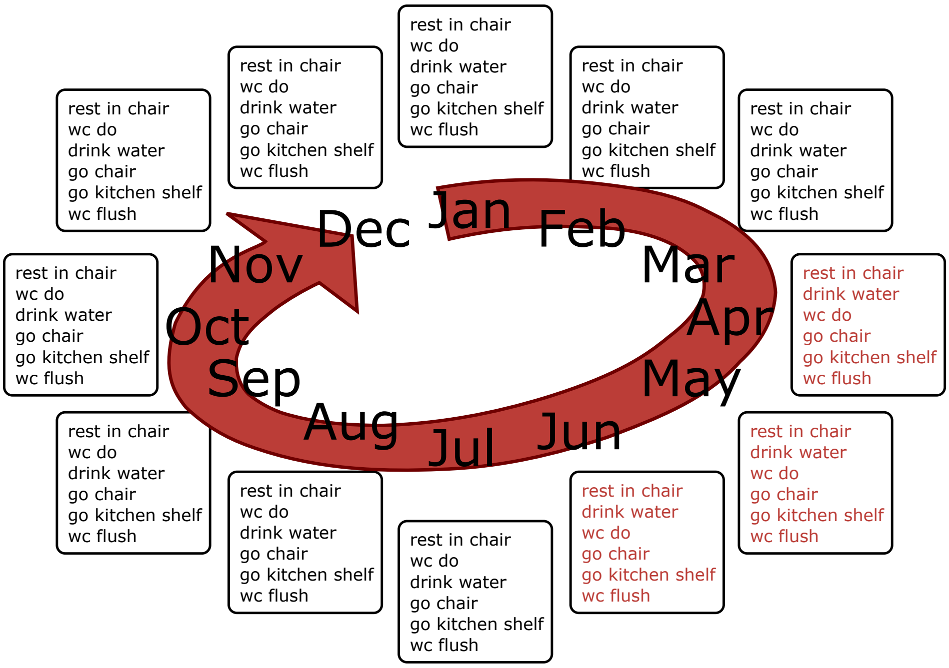


Figure 5.4: Variation in top activities of relaxation routine over 12 notional months. Activities are ordered by probability and changes in order are highlighted in red text.

a discovered routine is visualised for this routine, subjectively labelled as relaxing, in figure 5.4 by displaying the top 6 activities for each month for the whole year. The months with different orders of activities are highlighted in red. Routines 4 and 10, subjectively labelled as 'Eating and social' and 'Eating' demonstrate the largest variation in probability of the activities in the routines. The probability of the top activity relating to eating in each of these routines decreases in probability, although other activities in both routines that are related to eating increase in probability.

Analysing the raw simulated data confirmed that increasing the eating frequency parameter caused an increase of activities associated with eating. Furthermore, the most likely routine for each 30 minute time slice was determined from the results of the DTM. A matrix visualisation of the variation in frequency of each routine over twelve months is shown in figure 5.5. The colour corresponds to the normalised number of occurrences of the routine for each month. It can be seen that most of the routines either have very little variation (remain the same colour) or change frequency at random (jumping between pink and blue). However, it can be seen that the routine subjectively labelled as 'Eating' varies gradually from blue to pink from month 1 to 12, indicating that this routine becomes more frequent with time. The routine labelled 'Getting ready' also varies gradually, but decreases in frequency with time.

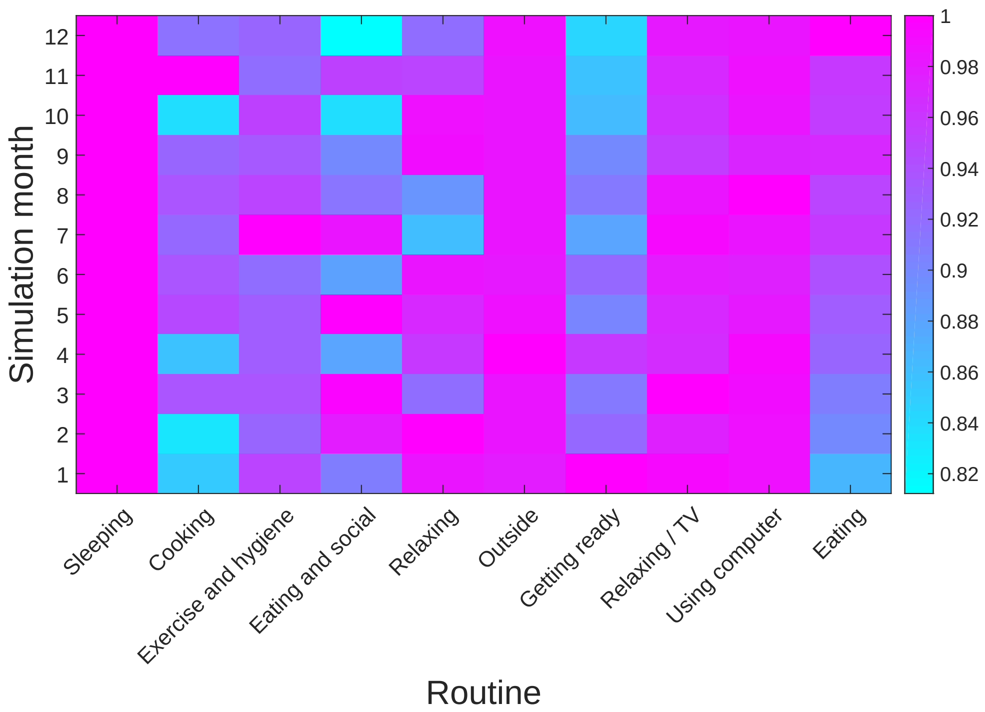


Figure 5.5: Visualisation of the normalised frequency of each routine detected by the DTM, over twelve months. Blue is lowest frequency and pink is highest.

5.2.2.2 Results of changing WC use frequency parameter

The results in this section are for dataset B where the WC use frequency parameter was changed, with all activities included listed at the low level of abstraction. For three DTMs estimated for K = 5, 10 and 15 routines, the top activities in each routine were visualised. For the 5 routine model, the ‘wc_do’ activity occurs in the top 6 activities for all months of two routines and for the last 3 months of another routine. In all of the routines, the probability of this activity only varies by 0.002 at most across the twelve months. The 10 routine model only has this activity in the top 6 activities of one of the routines and its probability is slowly decreasing by a total of 0.013 over the whole twelve months. Finally, for the 15 routine model there is one routine where the top four activities are related to using the toilet. The most likely activity is ‘wc_do’ and its probability decreases gradually by a total of 0.089 over twelve months. The other three related activities all increase in probability over time however.

The model with 15 routines contains one routine that is only related to using the toilet. Analysis of the frequency of occurrences of the most likely routine for each 30 minute time slice showed that this routine is never the most likely. Looking at the probabilities of the routine related to using the toilet for each time slice over the twelve months showed that overall there was a gradual increase with time. In contrast, the 10 routine model does not have a routine specifically related to using

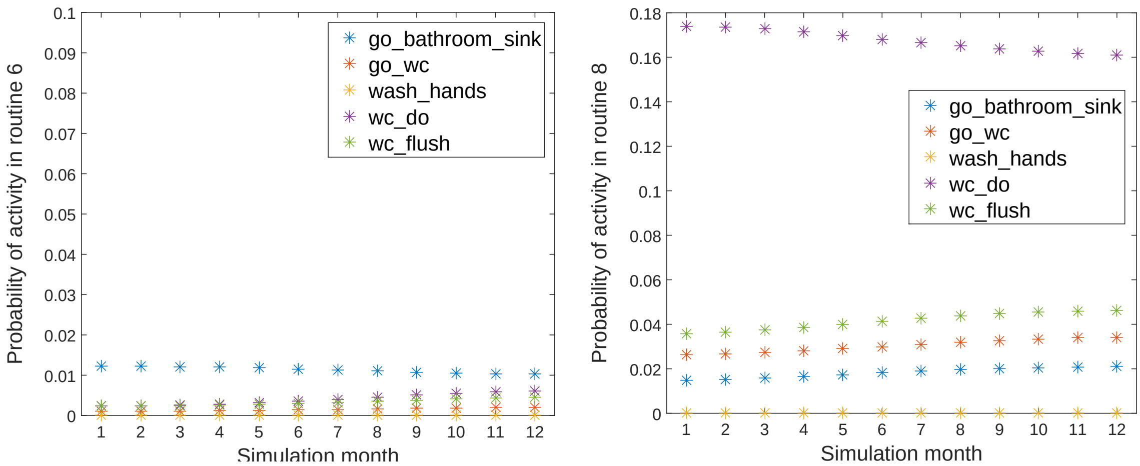


Figure 5.6: Variation over time of probability of activities related to using toilet in two discovered routines (6 and 8) for the 10 routine DTM estimated using dataset B.

the toilet and the only routine that has a relevant activity in the top 6 does not increase over time. Visualising the probabilities over twelve months of the toilet related activities for each routine shows that for two of the routines, there is a slight increase in the probability of some of the related activities, as shown in figure 5.6.

However, if it was not known that the WC use frequency parameter had been increased it would not be obvious to look at these activities specifically. If all of the activities are visualised at once then the changes in the toilet related activities do not stand out as changing significantly more than other activities by visual inspection. Therefore it is not possible to confidently identify changes from viewing the results in this way. The trends of changes in activities can be made clearer by highlighting only those activities in routines which demonstrate a change over twelve months of more than 25% of the minimum probability of the activity occurring in the routine, given that the minimum is greater than 0.001. The change in the activity over twelve months is then scaled to the range 0 to 1 in order that the direction of change can easily be seen, as shown for routines 6 and 8 in figure 5.7. Making direct comparisons between activities is not relevant as each one is individually normalised to the range 0 - 1 to allow clear observation of the trend of the change.

For the 10 routine model, the results of highlighting the activities that change the most relative to their minimum probability shows that such changes occur in 6 out of the 10 routines. Two of these routines, 6 and 8, have increases in toilet related activities, figure 5.7, as was found by specifically looking at these activities. However, two of the other routines, 1 and 9 also show strong trends of increasing or decreasing in activities that are related to getting a drink. The other two routines that have highlighted activities have more of an oscillatory pattern of change over the twelve months, rather than showing a strong increase or decrease.

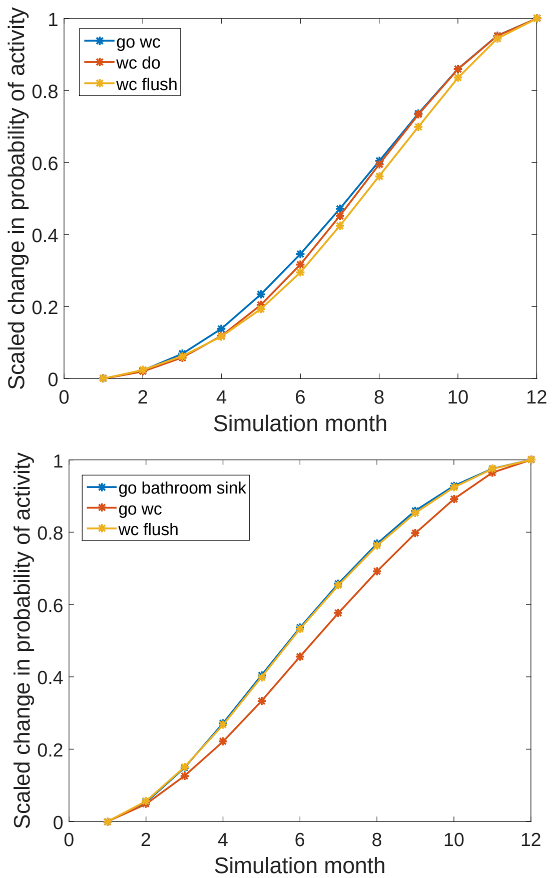


Figure 5.7: Changes in routines 6 and 8 that vary more than 25% of their minimum value, which is greater than 0.001. The range of the probability for each activity is individually scaled (between 0 and 1), hence comparisons across activities are not meaningful, however it does allow a comparison of trends.

As for the 10 routine model, the 5 routine model shows the probabilities of toilet related activities increasing within routines. For 4 out of 5 of the routines there are increases in probabilities over time of relevant activities. However, as for the 10 routine model, it would not be possible to visually identify these changes from others if they were not being specifically looked for. Highlighting all changing activities, that vary by at least 25% of their minimum probability, reveals that increases in toilet related activities can be found in 3 of the routines, as shown in figure 5.8. Increases in other activities co-occur in these routines though and one of the routines shows only a decrease in the 'wc-do' activity and changes in other non-related activities.

5.2.2.3 Results of changing activity data properties

The results in this section are for datasets C, D and E, where the WC use frequency was changed and the activity data properties were varied by removing long sleep activities and/or using the activities listed at a high level of abstraction. DTMs were

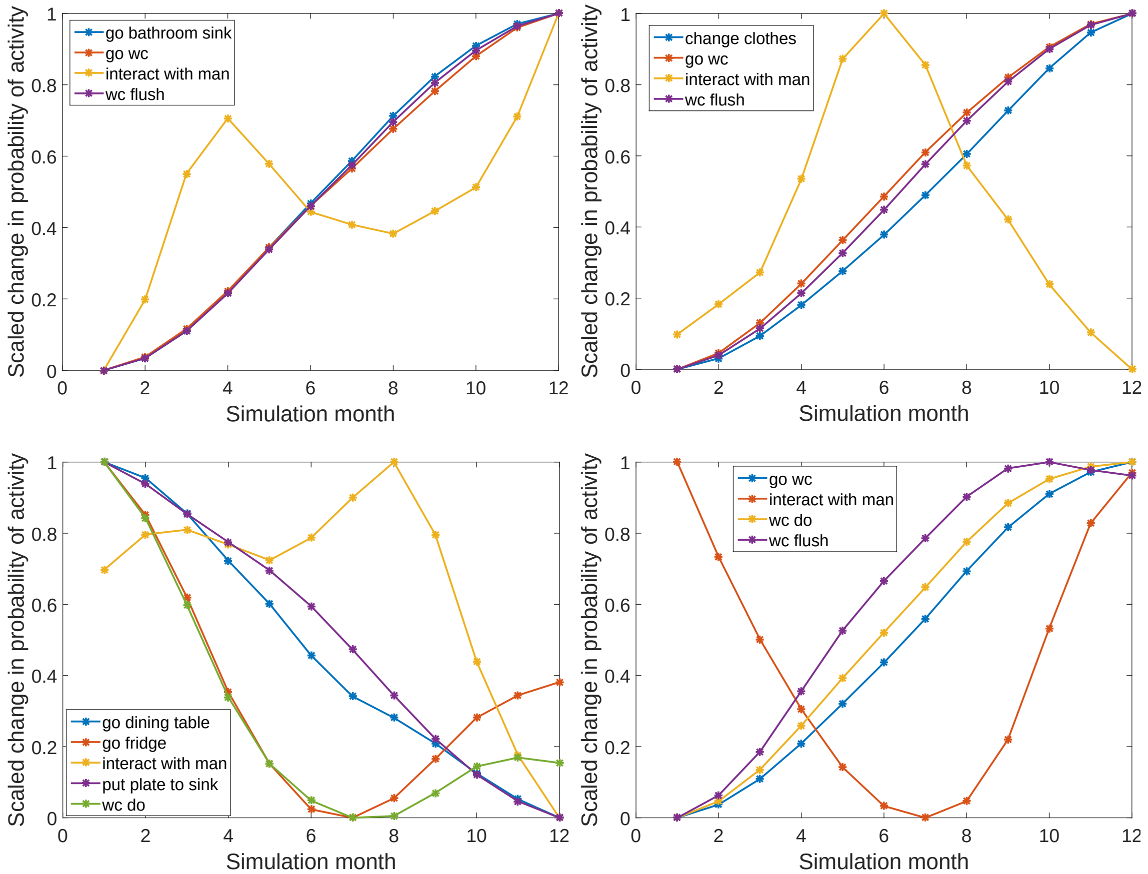


Figure 5.8: Highlighted changes individually scaled to between 0 and 1, that vary the most relative to their minimum probability, in four routines discovered by a 5 routine DTM estimated using dataset B.

only estimated for K = 5 and 10 for each dataset as using K = 15 produces routines that are too specific i.e. one routine for using the toilet.

For the 5 routine model using dataset C with long sleep activities removed, visualising the probability of the toilet related activities over time showed that four routines in this model had an increase in at least one relevant activity. These results are very similar to those for dataset B without sleep removed. Highlighting the changing activities in each routine showed that increases in the probability of toilet related activities were found in all of the routines except one. Three of these four routines also had other highlighted changes in unrelated activities which were oscillating rather than showing a steady increase or decrease in probability. One routine had increases in toilet related activities co-occurring with an increase in the 'drink water' activity. The final routine only had an oscillating change in one non-related activity.

A 10 routine model was also created for dataset C and the resulting probabilities of the relevant activities over time were visualised. The results showed that toilet related activities only occurred in one of the ten routines. Moreover, four of the top six activities in this routine are related to using the toilet. Determining the most likely routine for each 30 minute time slice shows that this routine is never the most likely. However, looking at the probability of the routine over the twelve months demonstrates that there is an increase with time.

When the vocabulary is changed to the activities at a high level of abstraction but sleep is not removed (dataset D), for the 5 routine model, the WC activity occurs in the top 6 activities of four routines. The probability of the activity is increasing in every routine, but sometimes only by a small amount. Highlighting changes in routines showed that the WC activity is increasing in two of the routines. However, the 'someone at entrance' activity is highlighted for four of the routines, although it oscillates in probability rather than showing a steady trend. The model with 10 routines only has the WC activity in the top 6 activities of one routine, with a relatively high probability. This routine is very rarely the most likely for each 30 minute time slice, but looking at the probabilities demonstrates an overall increase with time.

As the routines found using the K = 10 model for datasets C and D, with sleep removed or activities at a high level of abstraction were too specific, only a K = 5 model was estimated for dataset E with both of these properties. The results showed that the WC activity occurs in the top 6 activities of all the routines. Moreover, the probability of this activity is increasing in four of the routines and nearly stable in the other routine. Highlighting changes within routines revealed that the WC activity is increasing in three of the routines, however changes in other activities are also highlighted. In particular, the 'someone at entrance' activity occurs in all of the routines, although it has a varying probability in four of the routines and a decreasing probability in the other routine.

5.2.3 Discussion of detecting changes in routines

This section has investigated the use of dynamic topic models with simulated activity datasets to determine if it is possible to find a known change over time. In particular, the effect of the activity data properties on detecting changes was considered. A range of different visualisations were explored to determine the best method for automatically identifying relevant changes occurring at the activity level, i.e. the probability of activities within a routine changing over time.

Listing the top activities and their probabilities over time for the 10 routines found for dataset A, where the eating parameter was changed, showed that there was little variation for the majority of activities. This was as expected because all of the simulation parameters were kept constant, except for the eating frequency parameter. Viewing the results in this format provides a high level of detail but it is time consuming to consider all of the individual probabilities and draw useful information from the results. Therefore this is not a suitable method for presenting the data to clinicians. Figure 5.4 demonstrated how changes in the importance of activities within a discovered routine can be visualised in a summary diagram where changes in the orders of activities are specifically highlighted, helping to identify the changes quickly and easily. These changes can be linked to behaviour changes which help to provide users and clinicians with detailed data for aiding diagnosis.

The routines discovered did not show an obvious pattern of eating activities increasing in probability within routines. This suggests that the change did not occur at the activity level as expected. The matrix visualisation of the change in frequency of each routine over the twelve months confirmed that the change was occurring at the routine level. It was also noted that other routines showed a decrease in frequency. The probability of all the routines sums to one for each small time slice and therefore an increase in one routine can cause a decrease in others. In real terms this is equivalent to the idea that the number of hours in a day is fixed and hence if more time is spent on eating then less time must be spent on other activities. This impact may be seen as a decrease in one other routine or be spread across all other routines depending on the nature of the person and the change that is occurring.

The results for dataset B, where the WC use frequency parameter was changed demonstrated that the number of routines affects the level at which the change is seen. This is due to the fact that as the number of routines increases they become more specific i.e. only one or two activities have a high probability and the majority of activities have a probability of approximately 0. Therefore the routines are more like activities and hence changes will be seen at the routine level rather than the activity level. In the case of the 15 routine DTM for dataset B, one of the routines is only related to using the toilet and hence the change is seen at the routine level. However, because this routine is actually still short in duration, in comparison with the 30 minute length of the short time slices, it is never seen as the most likely routine and therefore the change is only seen if the probabilities of the routine are plotted over the full twelve months. Changes at this level can be of interest to clinicians but are not discussed further here as they are not the focus of this work.

When there are fewer routines each routine contains a larger mix of activities and hence is less specific. The results for the 5 and 10 routine models demonstrated that the change in WC use frequency can be seen at the activity level. This indicates that the models are sensitive to the relationship between the duration of the activities that are changing and the length of the routines that they occur in, which is affected by the choice of the number of routines and how specific they are. Although it is possible to find the WC use frequency change at the activity level, the results also showed that it is not the only change that was highlighted by considering activities that varied by over 25% of their minimum probability of at least 0.001. In particular, the 'interact with man' activity was highlighted in several routines but had an oscillatory pattern of change. The simulator used includes an element of randomness in determining the priority of each activity at any given time. This oscillatory change pattern could be attributed to this random change in the simulation, which is reflective of the complex nature of real life.

Other activities are also highlighted and show a trend of increasing or decreasing over the twelve months, such as 'change clothes'. These activities are not related to the parameter that was changed but can still be valid changes caused by the random element in the simulation or as an indirect result of the known change. Generally, these activities are isolated i.e. one activity is highlighted as changing but other associated activities are not highlighted, whereas the activities related to the known change appear in groups that change at the same rate. These additional changes reduce the confidence in the identification of the known change. Future work in collaboration with clinical experts, using real datasets need to evaluate the properties of the data and model that will produce clinically relevant information to help aid diagnosis.

Changing the properties of the dataset by removing long sleep activities and/or using a vocabulary of activities at a high level of abstraction affected the relationship between the number of routines chosen for the model and the level at which the change was seen. For datasets C, D and E the routines discovered by a 10 routine model were too specific and hence the change was seen at the routine level. Removing the long sleep activities means that there are fewer small time slices used as input to the model. Hence, each of the remaining small time slices will be a larger percentage of the whole dataset and so the model can find a better fit to this data as it is not dominated by time slices with only sleep. The resulting routines are more specific for the same parameter K, number of routines. Using the smaller vocabulary of high level activities means there are fewer activities listed in each time slice, hence the routines discovered have a smaller mix of activities and are more specific.

For the 5 routine models the change in the WC use frequency parameter could be seen at the activity level for these datasets (C, D and E). Other changes unrelated to the known change were also highlighted, similar to the results for dataset B. For example, the oscillatory pattern of the 'someone at entrance' activity when using activities at the high level of abstraction is equivalent to the behaviour of the 'interact

Table 5.5: Summary of changes found at the activity level, of more than 25% of the minimum probability of the activity, for models with different numbers of routines and dataset properties.

ID	Routines	Unrelated changes	Related changes	Details of related changes
B	5	0%	80%	60% increases, 20% decrease, all with unrelated changes
B	10	40%	20%	All increases, no unrelated changes
C	5	20%	80%	All increases with unrelated changes
D	5	40%	40%	All increases with unrelated changes
E	5	40%	60%	All increases with unrelated changes

with man' activity for the low level of abstraction.

Overall the known change of increasing the WC use frequency, was found at the activity level for datasets B to E using models with K = 5 routines and for dataset B using a model with K = 10 routines. For each of these models, table 5.5 summarises all of the highlighted changes found and whether they were related to the known change. The percentage of routines where all of the highlighted changes were unrelated to the simulation parameter that was varied are given. The percentage of routines where at least one of the highlighted changes was related to the known change is also presented. The details of the trend for related changes and whether they occurred with unrelated changes are listed to indicate how clearly they can be seen. When there is a larger percentage of routines with a related change and all of the related changes demonstrate the same clear trend this increases the confidence in the change found. The models estimated with 5 routines for datasets C and E, where long sleep activities were removed and the activities were at a low and high level of abstraction respectively, demonstrate the clearest visualisation of the known change at the activity level.

5.3 Detecting unknown changes in routines

The initial experiments suggested that changes occurring at the activity level can be identified more easily when estimating a model with 5 routines, using a dataset where long sleep activities have been removed. For these experiments the simulation parameter changed over time was known, introducing a bias when exploring the results. In a real world dataset the changes occurring will not be known a priori. This section describes an experiment where simulations were run without the author knowing the parameters that were changed, in order to remove this bias. Dynamic topic models were estimated for each dataset and the results were visualised to detect any changes, which were then compared with the revealed variation in simulation parameters.

5.3.1 Generating datasets with unknown changes over time

An unbiased third party ran three simulations (X, Y and Z) using the procedure described in section 5.1.2. For the human behaviour profiles a total of 6 options were provided, where all of the parameters were to be set to 50% except for one or two of the parameters. For each option the parameters to be increased from 25% to 80% by 5% each month are given in table 5.6. The instructions stated that two of options 1-4 and one of options 5 or 6 should be chosen; one of these for each simulation, in any order. The event logs generated for each simulation were given to the author without revealing the parameters selected.

Table 5.6: Options of parameters to vary from 25% to 80% by 5% each month, for simulations where the changes were unknown to the author

#	Parameters to change
1	Drink frequency
2	Eat frequency
3	TV frequency
4	Exercise frequency
5	Drink frequency and WC use frequency
6	Eat frequency and TV frequency

From the event logs for each of the three simulations a total of six datasets were created, as summarised in table 5.7. Long sleep activities were removed for all datasets and for each simulation there was one dataset with activities at the low level of abstraction and one at the high level. A 5 routine dynamic topic model was estimated for each of the six datasets and the vocabulary for each model consisted of all possi-

Table 5.7: Summary of properties for datasets generated from simulations with unknown parameter changed over time.

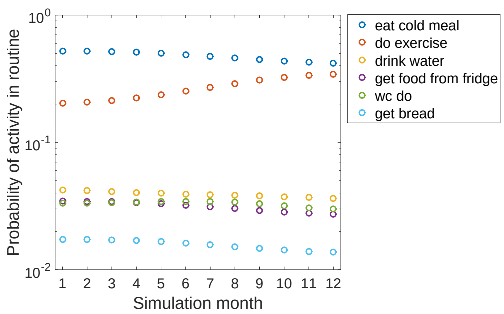
ID	Simulation	Long activities	Abstraction level
F	X	No	Low
G	X	No	High
H	Y	No	Low
I	Y	No	High
J	Z	No	Low
K	Z	No	High

ble activities that could be simulated at the relevant level of abstraction. As for the initial experiments, 30 minute time slices were created from 1Hz activity data with a sliding window of 2.5 minutes and each month was considered a large time slice.

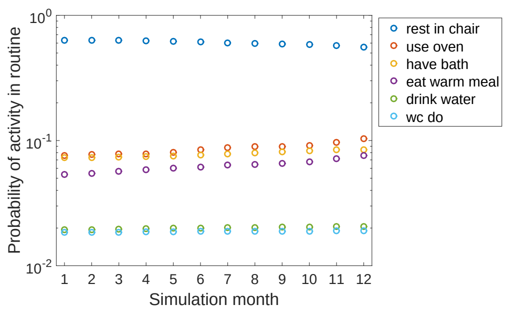
5.3.2 Changes in routines found using dynamic topic models

The results from models estimated for datasets F and G suggest that the exercise frequency simulation parameter was increased in simulation X, indicating option 4 was selected. The reasons for concluding this are discussed here. Visualising the probabilities over time of the top 6 activities in each routine, discovered by the model estimated using dataset F, shows an increase in the 'do exercise' activity in routine 1, as shown in figure 5.9a. A log scale is used in order to visualise all of the activities at once. There are also changes in activities relating to eating and using the toilet in two of the routines, therefore this alone is not enough evidence to draw a conclusion.

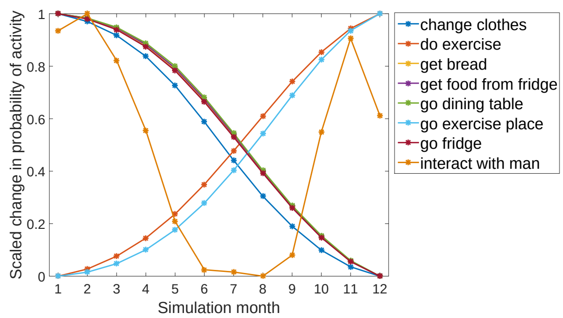
Viewing the highlighted changes, shows that every routine has the 'interact with man' activity, however it has an oscillatory pattern of change rather than showing



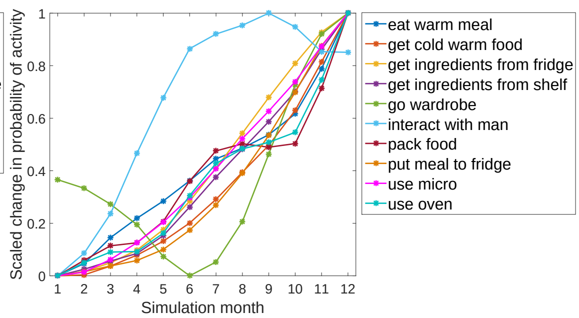
(a) Routine 1 top activities



(b) Routine 2 top activities



(c) Routine 1 scaled changes



(d) Routine 2 scaled changes

Figure 5.9: (a, b) Probabilities over time of top activities and (c, d) highlighted changes, in two routines estimated by a DTM using dataset F.

a strong increase or decrease. Therefore it is unlikely to be related to the changing simulation parameter but can still be a valid change occurring in the data due to the random element in the simulation. Routine 1 shows increases in highlighted changes for exercise related activities and decreases for eating related activities, as shown in figure 5.9c. Whereas, routine 2, figure 5.9d shows increases in many activities that are related to eating. This is also inconclusive as the trend of the highlighted change for eating related activities is opposite in the two routines.

As it is possible for changes to also occur at the routine level the most probable routine for each 30 minute time slice was determined. The variation in frequency for each routine across the full simulation duration is visualised as a matrix, as shown in figure 5.10. The colour corresponds to the normalised number of occurrences of the routine for each month, where blue is the lowest and pink is the highest frequency. It can be seen that routine 1 is occurring more often and routine 2 is decreasing in frequency.

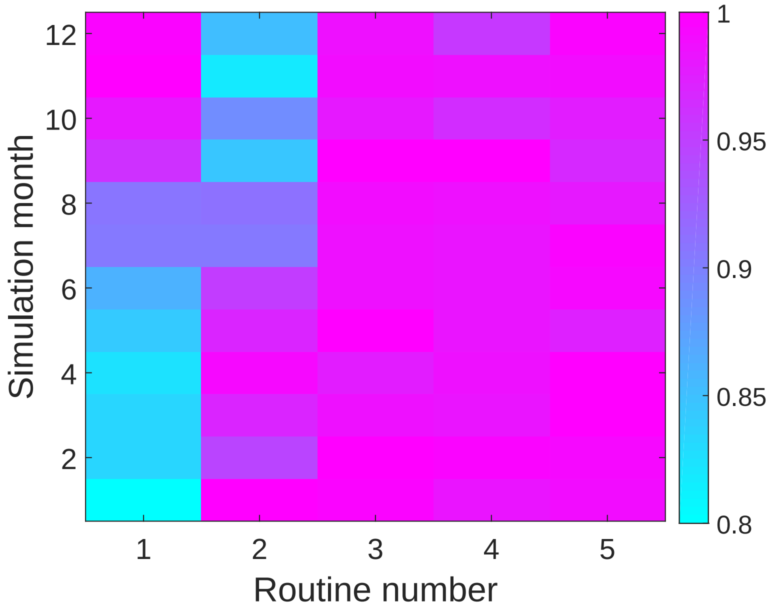


Figure 5.10: Visualisation of the normalised change in frequency of all routines over time. Blue is lowest frequency and pink is highest. Results from a 5 routine DTM using datset F.

Combining this information shows that exercise related activities are becoming more likely within routine 1, in addition to the routine increasing in frequency overall. In contrast, although eating related activities are increasing in probability in routine 2, overall the routine is occurring less frequently. Furthermore, the eating related activities in routine 1 are decreasing, although the routine is becoming more frequent. Hence the changes at activity and routine level effectively cancel each other out for the eating related activities, whereas for the exercise related activities they have a compound effect. This strongly suggests that the simulation parameter changed over time is the exercise frequency. The same patterns can be found in the results from the dynamic topic model estimated for dataset G.

For datasets H and I, the results from the dynamic topic models suggest that both the drink frequency and WC use frequency simulation parameters were increased, indicating option 5 was selected. The evidence leading to this conclusion is presented here. Visualising the top 6 activity probabilities over time for the model estimated using dataset H shows that there are changes in activities related to drinking and using the toilet in three of the routines. There are also changes in other activities across the routines but none of these are as consistent, for example the 'go tv' activity decreases in routine 2 but activities related to watching tv do not occur in the other 4 routines. This evidence alone is inconclusive.

Finding the highlighted changes for each routine gives further insight into the results. This shows that WC use related activities are increasing in 80% of the routines and drink related activities are increasing in 60% of the routines. The other highlighted changes that occur are much more variable and the probabilities tend to oscillate over time. Moreover, for this dataset, it was found that there was very little variation at the routine level, indicating the change is occurring at the activity level as desired.

The results of the model estimated using dataset I show that when visualising the top 6 activities in each routine over time the WC activity is changing in 100% of the discovered routines and the drink activity is changing in 60%. There are also changes in other activities across all of the routines but they generally show an oscillatory pattern of change. A similar picture can be seen in the highlighted changes, as shown in figure 5.11. For three of the routines the highlighted changes show increases in WC, drink or both. However, the WC activity is decreasing in two routines. The other highlighted changes all demonstrate an oscillatory pattern of change. Overall the only consistent changes with a strong trend are for activities related to drinking and using the toilet, suggesting these are the simulation parameters changed over time.

The results of the dynamic topic models estimated using datasets J and K suggest that the eating frequency simulation parameter was increased, indicating option 2 was selected. Considering the probabilities of the top 6 activities in each routine, found by the model estimated using dataset J, suggests there is no obvious change at the activity level. Routines 1 and 4 show a decrease in activities related to using the toilet and routine 2 shows a decrease in the 'drink water' activity. Routine 5 shows the most changes, with eating related activities both increasing and decreasing and the 'wc do' and 'drink water' activities increasing. The highlighted changes

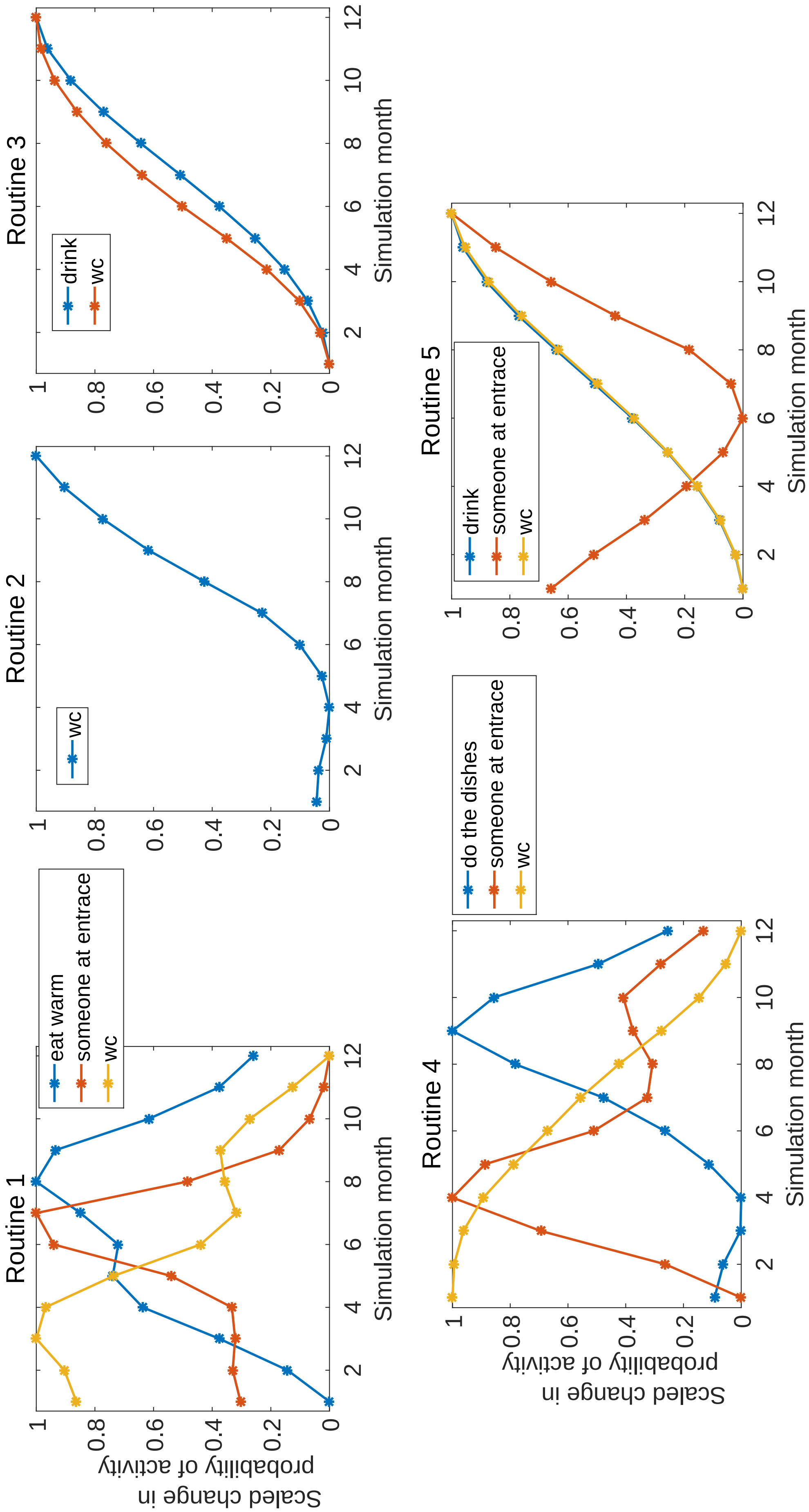


Figure 5.11: Highlighted changes for all routines for a 5 routine DTM estimated using dataset I.

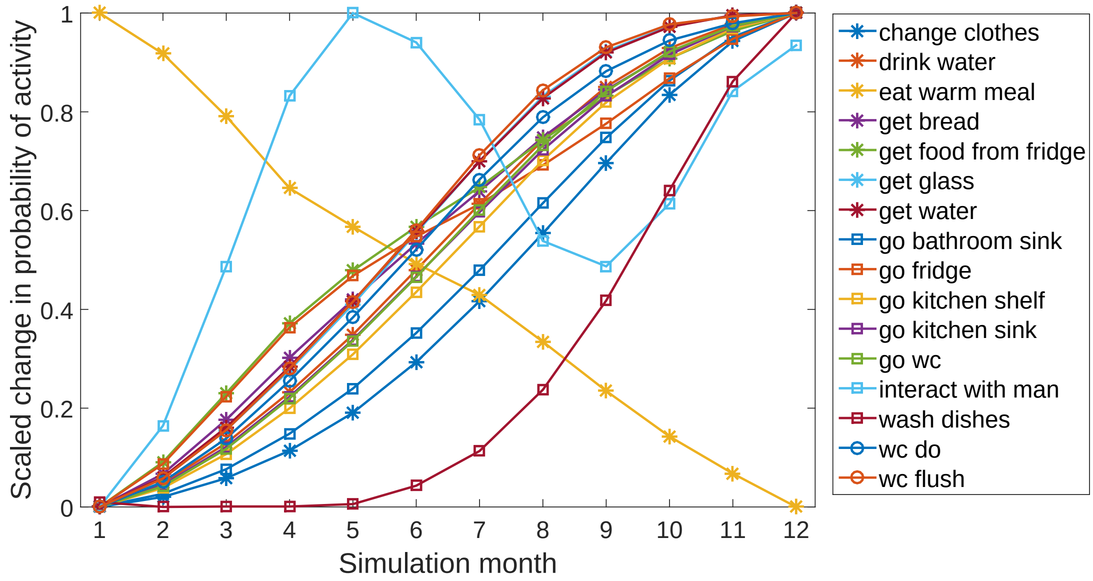


Figure 5.12: Routine 5 highlighted changes for DTM estimated using dataset J.

also demonstrate no obvious consistent change. In particular, routine 5 has many highlighted changes, as shown in figure 5.12, including several activities related to eating.

The matrix representation of the variation over time of the frequency of each routine, figure 5.13, demonstrates that routine 5 is occurring more often over the twelve months. This shows that the change is happening at the routine level. Therefore, although the eat warm meal activity is decreasing within routine 5, the increase in several other eating activities and the overall increase in the routine provides strong evidence that the change is in the eating parameter. The results of the model esti-

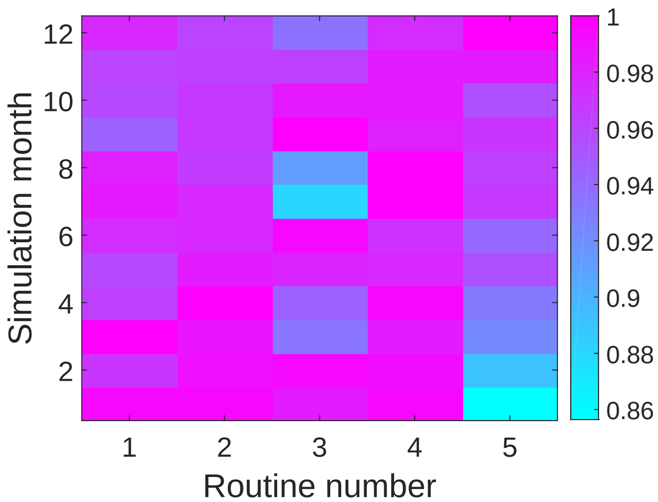


Figure 5.13: Visualisation of the normalised change in frequency of all routines over time. Blue is lowest frequency and pink is highest. Results from a 5 routine DTM estimated using dataset J.

mated using dataset K also do not demonstrate any obvious consistent change at the activity level. There is an increase over time at the routine level of routine 1, which has the top activity ‘eat cold’. In addition, the highlighted changes for this routine show the ‘do the dishes’ activity increasing.

After the results were analysed and the predicted options determined the original selections were revealed, as summarised in table 5.8. It can be seen that the correct option was successfully identified using the routines discovered by the dynamic topic models for every simulation. This demonstrates that it is possible for unknown changes over time to be identified using DTMs on a simulated dataset. However, some of the changes found were at the routine level rather than the activity level. Furthermore, additional changes were identified as well as those related to simulation parameters changed. The limited list of possible changes to be found means that there is still an element of bias and further work is required to validate this approach using long term annotated real world datasets.

Table 5.8: Results of detecting unknown changes in routines for three simulations with different parameters varied over time.

Simulation	Predicted	Actual
X	Option 4	Option 4
Y	Option 5	Option 5
Z	Option 2	Option 2

5.4 Chapter summary

This chapter has considered the use of dynamic topic models to identify changes in the probabilities of activities within discovered routines. The initial experiments demonstrated that these types of changes can successfully be identified for a simulated dataset with a known change. However, some of the changes occurred at the routine level, that is the probabilities of the activities within a routine remained similar over time but the routine occurred more frequently over time. It was found that the relationship between the duration of the activities in the dataset used as input and the number of routines specified affect how easily a change can be identified. The best performing combinations for the simulated datasets investigated were identified as models with 5 routines and removal of long sleep activities from the data. The vocabulary used did not have much effect on the performance.

These selections of activity data properties and model parameters were used to estimate DTMs and identify unknown changes in six datasets created from three different simulations. The results showed that the DTMs can successfully be used to identify changes when they are not previously known. This indicates that DTMs have potential for identifying changes in routines over time occurring in real activity data. There are limitations with these results as they are for simulated data with a maximum of two parameters changed over time and a limited number of options to choose from when identifying the unknown changes.

Different visualisations and techniques to explore the routines discovered by DTMs have been explored. The results demonstrated that considering how the probabilities of the top activities in each routine vary over time can reveal interesting patterns. However, interacting with the results at this level is time consuming and would not be suitable for end users unless summary diagrams that clearly outline the changes can be automatically generated. Presenting all of this information, including the magnitude or significance of a change, concisely and clearly is a challenging data visualisation task.

A new visualisation technique was proposed that highlights changes in activities that meet given criteria. For the simulated datasets the requirements were that the activity must vary by over 25% of its minimum probability and that the minimum probability must be at least 0.001. This approach uses the relative change of the probability rather than the absolute change. The probability of an activity in a routine is the likelihood of that activity if the routine is occurring and hence activities of shorter duration will have lower probabilities. This does not mean that a small change in absolute terms is not significant if it has a large change relative to its initial probability. The highlighted changes are scaled from 0 to 1 in order to visualise the trend of the change clearly. A change that oscillates over time is less likely to be significant if it is related to natural phenomenon, such as seasonal changes or random fluctuations. These criteria were shown to be successful in highlighting the large relative changes in the simulated datasets considered. However, they may not relate to significant real world healthcare criteria for a real dataset.

Future work needs to address the limitations of the existing results. Firstly, a larger experiment on more complex simulated datasets with multiple parameters changing and no limited list of permutations is required. This would demonstrate the ability of DTMs to detect changes in data that are more similar to a real-world dataset. A large number of varying parameters may cause the results to highlight too many changes to be helpful in understanding the data. In addition, the activity labels generated by the simulator are assumed to be a perfect representation of the person's behaviour. In reality, activity recognition algorithms using raw sensor data will not be completely accurate and hence noise will be introduced into the system. The impact of this can be investigated using simulated datasets by manually varying deletion and insertion errors of the generated activities, as done for LDA and other topic model variants by Seiter et al. [125]. Secondly, a real-world long term dataset with known changes would be needed to demonstrate that the method can work on data that is not simulated. Furthermore, valid criteria for detecting significant changes related to health and well-being will need to be established in collaboration with relevant experts.

The combination of activity data properties and model parameters for this experiment were chosen in relation to the data generated by the simulator. Only some of the changes made were discovered at the activity level. This is due to the variation in duration of activities. A new vocabulary where activities are selected to ensure that their mean duration is within a specified range would improve the performance of the DTM. The length of the small time slices would also be set in relation to the length of activities so that most of the time slices contain a mixture of different activities. These decisions should be made in conjunction with expert clinical opinions in order to be able to identify changes that are significant in terms of healthcare. Different models may be required for different situations if changes are occurring at varying levels of granularity.

Certain changes in behaviour can be detected simply by monitoring how frequently relevant activities occur. However, if there are a lot of activities being recorded then it would be time consuming to review and identify changes at this level. It is quicker for a clinician to be given a summary at a higher level, only reviewing more detailed data if necessary. The benefit of using DTMs to identify changes within routines is that this processing can be done automatically to provide the user a concise summary. Only analysing changes at a higher level directly would risk the loss of important information that is hidden in the data. DTMs enable both a high level summary whilst still incorporating useful information from the more detailed data.

In conclusion, dynamic topic models are useful for detecting changes over time in the probabilities of the activities that compose a routine. DTMs should be used carefully as certain conditions and assumptions need to be met for them to perform well. This approach is not a one size fits all solution as it is affected by the model and dataset parameters and how they relate to each other. For example, a model that performs well for detecting changes in activities that only last a few minutes within a routine that is varying on a weekly basis will not perform well if applied to a dataset containing changes in activities that last half an hour within a routine that varies every few months. This is a limitation of the approach as it is not generalisable and needs to be tuned using expert knowledge and contextual information for different scenarios.

Chapter 6

Detecting Eating Event Types in Nutrition Data

Many chronic conditions are related to diet, such as diabetes and obesity [7, 16]. Therefore, one important focus for a healthcare residential monitoring system is the consumption of food and drink. Recently, nutrition research has moved away from the traditional approach of investigating relationships between single nutrients and specific health outcomes to take a more holistic view by considering dietary patterns [184, 223]. This approach reflects the fact that links between diet and health are multifaceted and complex. Furthermore, it acknowledges that single nutrients are not consumed alone; in fact people eat items of food, which contain multiple nutrients. Equally, meals or snacks represent combinations of food consumed together within an eating event. The ingestion of nutrients and foods in combinations can interact with each other and therefore considering how they occur together could help to reveal stronger relationships to health outcomes than considering them individually [16,223].

Several studies consider the association between different eating behaviours, especially snacking, and the impact on diet quality and health. Results throughout the literature are often conflicting and this is, in part, due to not having consistent definitions for different eating events, including what a snack is [17,18,184]. As highlighted in section 3.3.3, there are a variety of methods used in studies to define eating events; moreover, these definitions are often biased by cultural norms [184]. Clear definitions of eating events are required to enable consistent research into related health outcomes and determinants of eating behaviours [18]. They would also inform public health guidance, to ensure that advice is relevant and useful. This chapter will consider how topic models can be applied to nutrition data to reveal patterns in the combinations of food groups consumed together in different eating events.

6.1 Applying topic models to nutrition data

The aim of applying topic models to nutrition data taken from food diaries, which has not previously been considered, is to identify combinations of food groups consumed together in an eating event. It is hypothesised that the results of the topic model will be representative of eating events composed of similar combinations of foods, which can be subjectively labelled as meals or snacks. This could provide a standardised approach for categorising eating events that could be used within the nutrition research community. In order to apply LDA to food diary data it is necessary to define the probabilistic model, as presented in 2.2, in the context of this application. This is not a trivial problem and this section describes the approach taken to find a suitable vocabulary and method for document creation for the National Diet and Nutrition Survey Rolling Programme (NDNS RP) dataset. This dataset includes dietary data from four-day unweighed food diaries, with each item consumed by a participant listed separately.

6.1.1 National Diet and Nutrition Survey (RP) dataset

The NDNS RP was a continuous cross sectional survey conducted from 2008 to 2012 [224]. The survey investigated food consumption, nutrient intake and nutritional status of people aged 18 months or older and living in private households, from a representative sample across the UK population. The sample was drawn using a multi-stage stratified random sampling procedure. Addresses were selected randomly from primary sampling units, which are clusters of addresses from the postcode address file for the UK. The aim was to collect data for 500 adults and 500 children each year, therefore for some addresses only children were selected to participate. Extra addresses for Wales, Scotland and Northern Ireland were selected to boost sample sizes for these countries to enable comparisons between countries, however these were not included in the available datasets (Core Sample Data) at the time of this work. In total there were 4,156 fully productive participants for whom a complete set of data is available. Full details of how the survey was conducted and all of the data collected is available from the UK dataservice [225]. This data was accessed under usage number 87742 on 23/02/2015.

The study includes detailed dietary data from unweighed four-day food diaries recorded by the participants, although participants who only completed three days were still classified as fully productive. The food diaries were completed over four consecutive days. Across the whole study all days of the week were represented as equally as possible, accounting for the preferences of the participants [224]. The information from these diaries were coded by trained individuals using the Department of Health's NDNS RP Nutrient Databank. Each item, or component part of composite items such as a sandwich, recorded in the diaries was assigned a food code and portion code [226]. The aim of coding food diary data is to establish the nutrient composition as accurately as possible. Recipes for composite foods can vary widely and hence they are separated into their constituent ingredients to improve accuracy. Variation between recording composite items in their entirety or as component parts depends on the availability of information supplied by the participant and data on food composition in the database.

The data from the entire survey is provided at individual and household levels and food diary data at food, day and person levels. For this work, the food level dietary data is generally used, which consists of 384,028 records, one for each food item recorded for each participant, on each day. This provides very detailed information about each item consumed, including the food name, food group, food subgroup, recipe group, nutrients, weight and energy content. In addition, contextual data is included, such as unique ID, age, sex, location eaten, company when eating, day of week and time of consumption. However, for some of the exploratory work the day level dietary data for foods was also considered. This consists of 16,540 records, one for every diary day for each participant, giving a record of the total weight of each food group consumed during the day. It should be noted that the food groups used in this section of the data vary slightly from those used in the food level data. Full details can be found in the NDNS RP documentation [182, 227] and are listed for reference in appendix B.1.

6.1.2 Choosing vocabulary

For the purpose of this work, an eating event is defined as all items consumed at a specific time on one day, as recorded to the nearest second in the food diary [228]. In other words, every entry in the food diary (a group of items consumed at a specified time) is equivalent to an eating event. The important information about analysing an eating event is its energy content, weight, constituent food items and time of day. Therefore, it was initially proposed that the vocabulary used for the topic model should include these aspects. It was thought the food names given in the dataset could be used directly as these are already mapped to food groups and subgroups so no information would be lost. To have a separate vocabulary entry for every weight, energy value and time would result in a verbose vocabulary e.g. a 40g (94 kcal) slice of white bread eaten at 7am would be one vocabulary item. Therefore the dataset was analysed to decide on realistic groupings which could be used as vocabulary items, e.g. weights between 1 and 50g could be one vocabulary item and energies between 0 and 100 kcal could be another.

A statistical summary of the dataset was used to determine suitable ranges for grouping the meal time, energy and weight variables. It was found that meal times recorded in the dataset ranged across the full 24 hour period. The maximum interval between any two meal times recorded for any participant across the whole dataset was found to be 10 minutes. The sum of the energy of all items in each eating event was also analysed. The majority of eating events had a total energy content of less than 2000kcal. Items over 2000 kcal were for example, 1kg of pizza or 500g of chocolate. A histogram of the total energy of eating events was created, automatically selecting the bin width to fit the data and reveal the shape of the underlying distribution. An optimal bin width of 20 kcal was found, as shown in figure 6.1, energies above 1000 kcal are not displayed here as the frequencies are relatively small. The same approach was taken for analysing the sum of the weight of all items in each eating event. It was found that the majority of eating events had a total weight of under 2kg. Typically beer and cider consumption contribute to entries over 2kg. The histogram of weights finds an optimal bin width of 10g and the distribution of weight up to 1kg can be seen in figure 6.1.

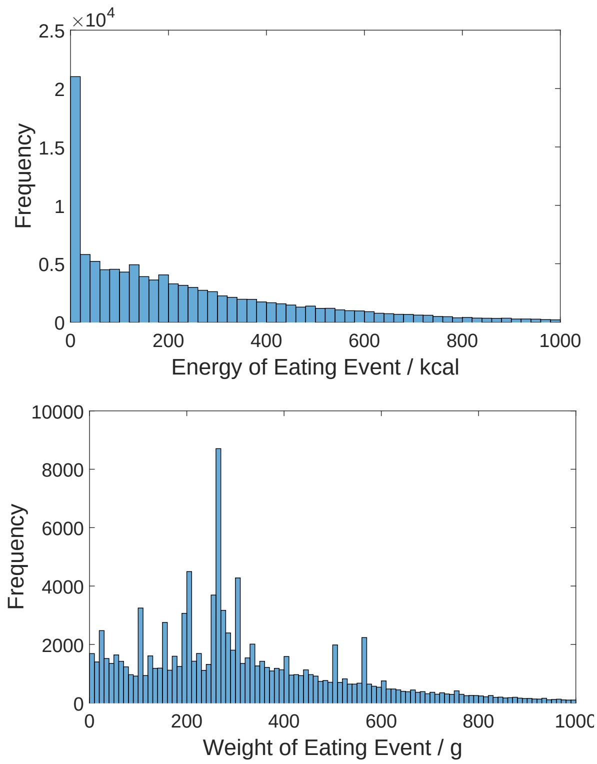


Figure 6.1: Histograms of sum of energy content and weight of all eating events in the NDNS RP dataset. The bin sizes are 20 kcal and 10g respectively.

Based on this analysis an initial set of vocabulary items were chosen as described below. For the energy, weight and time, each item was created from a prefix, representing the relevant type of data, and the associated range. This creates descriptive words that facilitate subjective analysis of the final results of a topic model. In total there were 4521 vocabulary items, each assigned a unique ID.

The initial vocabulary items were as follows:

- Food names of all items in dataset
- Meal Time ranges in 10 min intervals from midnight to midnight (MT00:00-00:10 to MT23:50-00:00)
- Energy content ranges in 20kcal intervals from 0 to 2000 kcal (E0-20 to E1980-2000) and energy content over 2000 kcal (E2000+)
- Weight ranges in 10g intervals from 0 to 2000g (W0-10 to W1990-2000) and weights over 2000g (W2000+)

Using this vocabulary with the assumption that a document in the original LDA model is equivalent of one eating event for this application restricts every vocabulary item to appear once at most in any document. If there was a larger quantity of any one item it would just be listed in the original diary data with a higher quantity. Only having items appear once per document would result in very sparse documents. One solution considered was to change the vocabulary to use the 60 food groups that exist in the dataset, rather than the individual food names, as each food group contains many food items and hence may occur more than once in a document. Other solutions based on changing the document structure were also considered and these are outlined in more detail in section 6.1.3.

Furthermore, having separate vocabulary items for energies, weights and times meant that the intrinsic relationship between a particular food item and its properties was lost. One option considered was to combine food names or food groups together with the weight, energy or energy density, for example, 'Bacon and Ham E80-100' could represent a 90 kcal portion of bacon. This would create unique vocabulary items that maintain the link between the properties and the item. However, this would lead to a verbose vocabulary where words would occur infrequently in documents, hence this approach was not considered appropriate for the modelling assumptions. Alternative solutions are described in more detail in section 6.1.3.

The separate weight, energy and time vocabulary items meant that some of the topics found by the model did not contain any food items or food groups in the list of most significant words. An example of a topic with only time and energy values in the top 10 words is given in figure 6.2b. Although such topics could still reveal interesting information about patterns between weights, energies and times; for investigating types of eating events it is desired to know about the specific food groups

(a) Main Meal Topic

(b) Non-food Topic

Figure 6.2: Examples of the top 10 food groups in two topics. Discovered by a model with 10 topics; a vocabulary based on food groups, weights, energies and times; and documents created per eating event. Displayed with font size proportional to the probability of the vocabulary item in the topic. The labels 'Main Meal' and 'Non-food' are subjectively assigned to the topics by the researcher based on the food groups listed.

involved. In contrast, figure 6.2a depicts a topic that is semantically coherent and can be subjectively labelled as being a 'Main Meal' based on the top 10 vocabulary items in the topic.

6.1.3 Approaches to creating food documents

For the original application of LDA to a corpus, a document is implicit and does not need to be defined. However, when applying the model to alternative types of data it is necessary to consider and define what is meant by a document. A document is represented as a bag of N words taken from the vocabulary but the boundaries of the collection are not defined for this application. For continuous data, a given time slice can be used to define a document, such as a 30 minute window, as used for the activity data in chapter 4. In contrast to the activity datasets, the NDNS RP dataset contains data from a large number of participants, rather than long term records for a few participants. Furthermore, food diaries are recorded as discrete eating events, at specific times, with varying intervals in between events for each participant. Therefore, using windows of a specified length for eating events recorded for one participant does not guarantee that all documents will contain data.

Initially, a document was considered to be a single eating event, as defined in section 6.1.2. Documents were created by recording all of the relevant vocabulary items for the eating event, from the original vocabulary of 4521 items. For weights, energies and times the vocabulary item representing the range in which the actual value occurs was used. As mentioned in section 6.1.2, creating documents based on a single eating event means that they are very sparse. Although using food groups is an improvement over using the food item names directly the documents remain sparse.

In order to reduce the problem with sparsity in the documents and to encode the relationship between a food item and its properties more explicitly, a new approach was taken to creating documents. Each food group was repeated in proportion to its energy content, using a vocabulary of food groups only. Using this approach the topic model discovered some semantically coherent topics which are representative of typical eating events. However, some of the topics were combinations of food groups that did not relate to a typical eating event when evaluated qualitatively. For example, a topic with the following top vocabulary items and associated probabilities in brackets, 'biscuits (0.59)', 'eggs and egg dishes (0.24)' and 'white fish coated or fried (0.16)' would be challenging to apply a subjective eating event type label to.

Repeating the food groups solves the sparsity problem, however if a topic is considered to be a type of eating event and a document is created based on a single eating event, then by definition, a document will only exhibit one topic. This does not support the assumption of LDA that a document will exhibit multiple topics. Therefore, the structure of a document was revised so that each document is created using the data from one full day. As described in section 6.1.1, the dataset also includes data at the day level, recording the total weight in grams, for each of the 67 food groups, consumed during the day. This data subset was used to create documents at a day level, with a vocabulary of 67 food groups, where each food group is repeated in proportion to the weight consumed on that day. The resulting topics from this approach were subjectively identified by a nutritionist expert to be more reflective of different lifestyle dietary patterns, rather than types of eating events. For example, the topic

Figure 6.3: Example of a topic, created from day level food group data repeated in proportion to weight of consumption, that is representative of a low-energy dense, low-fat, high-fibre diet. Displayed with font size proportional to the probability of the vocabulary item in the topic.

displayed in figure 6.3 is representative of food groups that would be likely to appear in a low-energy dense, low-fat, high-fibre diet [229].

As day level data did not produce the desired results, an alternative approach to creating documents with more than one eating event was taken. It was proposed to create documents using all eating events that occur in the dataset at a given time, t, to the nearest second, regardless of the person or day on which the event was recorded. Creating documents using this method means that the time of eating is still represented but documents can exhibit multiple types of eating events. This approach to creating documents was used together with various vocabularies, including food names and food groups as well as different representations, such as single items or repetition proportional to energy content.

Table 6.1: Summary of combinations of document structures and vocabularies investigated. Comments on limitations and qualitative evaluation of the topics discovered

Document structure	**No. docs**	**Vocabulary**	**Comments**
Single eating event	111031	food names (+ weight, energy, time) food groups (+ weight, energy, time)	Food names only appear once at most. Documents are sparse. Loss of connection between food item and properties. Topics with only weight, energy or time items.
Single eating event with items repeated proportional to energy	111031	Food groups	Only some topics are coherent
All food for one day with items repeated proportional to weight	16539	Food groups (from day level data)	Topics are reflective of dietary patterns rather than eating event types
All events at time t	1038	Food names Food groups	Topics with food names are too specific. Topics with food groups are semantically coherent.
All events at time t with items repeated proportional to energy	1038	Food names Food groups	Topics not semantically coherent. Vocabulary items have high frequencies, especially for food groups.
All events with energy in given range	401	Food groups	Documents too general, most contain all vocabulary items. Topics not semantically coherent.

Another approach proposed for creating documents representing multiple eating events was to use all eating events that occur in the dataset with a total energy content within a given range. The 20kcal ranges used for the original vocabulary and based on the analysis of the dataset are not suitable as this would only create 101 documents. Therefore, 5kcal ranges were selected to give 401 documents: one for each 5 kcal interval from 0 to 2000kcal and one for 2000kcal and over. A vocabulary of food groups was used with this document structure. The topics discovered had very low semantic quality and had no clear link with different types of eating events. It was noted that many of the documents contained the large majority of vocabulary items, making it harder for the model to distinguish between different documents and find relevant patterns.

6.1.4 Selected document structure and vocabulary

The limitations and qualitative evaluation of the different combinations of document structures and vocabularies investigated are summarised in table 6.1. This demonstrates that the most promising combination is a vocabulary of food groups in the NDNS RP dataset and a document structure of all eating events at time t. This combination matches the underlying assumptions of LDA, in particular, ensuring that documents exhibit multiple topics. In addition, the discovered topics are semantically coherent in the context of discovering eating event types. This document structure creates 1038 documents, one for each unique time recorded to the nearest second, that appears in the dataset spanning a full 24 hour period in total. For each document, the corresponding food group for every food item, in each eating event recorded at time t, is listed.

The selected document structure and vocabulary is used as the basis for the re-

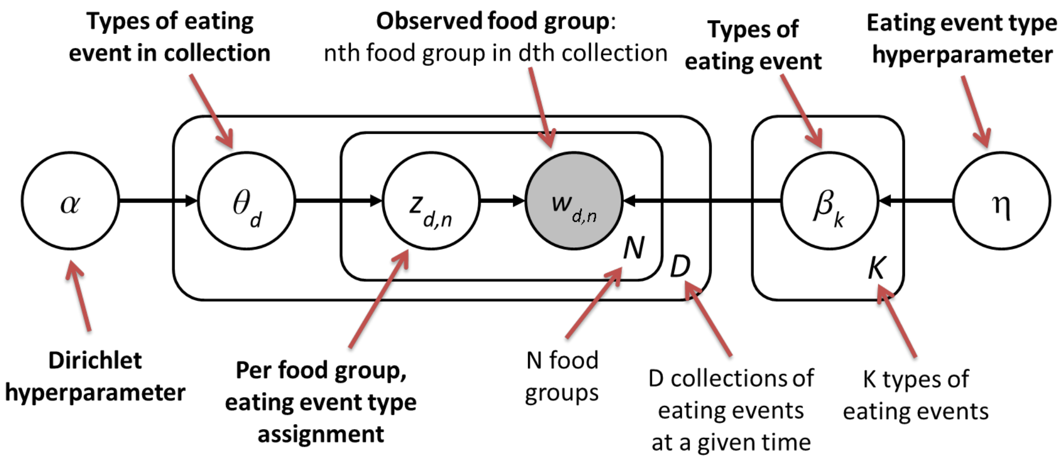


Figure 6.4: Applying LDA model to nutrition data.

mainder of the work presented in this thesis. The exact choice of food groups included in the vocabulary varies as required, the details of this are given in each relevant section. The representation of the LDA model, in the context of nutrition data for investigating types of eating events is given in figure 6.4. Here types of eating events are equivalent of topics; a collection of eating events at a given time are equivalent of documents and food groups are equivalent of words.

6.2 Detecting types of eating events

The selected document structure and vocabulary were used to carry out experiments to evaluate the performance of topic models in the context of understanding how foods are combined together in different types of eating events. A sensitivity analysis of the impact of different properties of the dataset was also conducted. Different methods for analysing the results of a topic model to assess how well it is performing were presented in section 2.4. This included both qualitative and quantitative techniques for evaluating the model in the given context. For this application, qualitative evaluation of the resulting eating event types is extremely important as these are what will be utilised by the nutrition research community to facilitate further research. Even if a model statistically performs very well, if the results do not have any tangible meaning then it is not possible to use them as a basis for future analysis. However, quantitative analysis still plays an important role, to ensure that the results are representative of the data. In addition, quantitative techniques can help confirm the selection of model parameters.

6.2.1 Experimental methods for investigating the performance of topic models applied to nutrition data

For consistency, the same C implementation of variational inference for LDA from David Blei [51] was used, as described in section 4.1.1. Although the data source is different, documents are still represented as sparse vectors of word counts in the same format as previously used. The preprocessing to generate the vocabulary list and create the documents with the selected structure from the NDNS RP dataset was done in MATLAB. Eating events were identified in the dataset by grouping the food level data on the 'ID', 'meal time' and 'day no.' variables. For the chosen document structure, every unique time in the full dataset was identified and for each of these, all of the corresponding eating events were selected. The food groups of all items consumed during these eating events were counted and used to create the sparse word vector for the document. The numbering system used to identify food groups in the dataset was used as the numerical representation of the vocabulary.[1]

For each model, the number of eating event types and the initial value of the Dirichlet prior, α, were set to the desired values, as specified in section 6.2.2. The default settings for the maximum number of iterations and convergence criteria were used, as done in section 4.1.1. The value of α was set to be estimated as part of the model so that the best fit to the data was found and a random initialisation was used for the eating event types distributions.

6.2.1.1 Qualitative and quantitative analysis for full NDNS RP dataset

Post-processing of the results was also carried out in MATLAB. The most intuitive way to qualitatively evaluate the results is to investigate the eating event types, which are the probability distributions over the vocabulary, β_k. The top food groups, with a probability greater than or equal to 0.02 [32], for each eating event type were determined. The list of food groups enables an assessment of how well the content of the discovered eating event types reflect reality. In addition, it is useful to be able to visualise these results in a more compact form for a quick evaluation. A tool was developed in MATLAB to print out up to the top 10 food groups of an eating event type in order of probability where the font size of the food group is directly proportional to the probability that it occurs in the eating event type.

Subjective assessment based on knowledge of different types of eating events was used to determine the semantic quality of the eating event types found and to apply a label that summarised the content of the eating event type for further analysis. For example, an eating event type with top food groups such as bread, margarine, ham, salad and similar, appears to be a combination of items that could form a sandwich or other typical lunch time eating event. This assessment was verified by a nutritionist expert, with experience looking at different combinations of food groups in the context of eating events. These subjective labels are used to aid visualisations of the results and as the basis of further analysis, as presented in chapter 7. However, it should be understood that they are only a representative description of the underlying probability distribution and are open to interpretation by experts in the nutrition research community as required. In essence the discovered eating event types represent the most likely combinations of food groups consumed by a representative sample of the UK population.

[1] In the NDNS RP dataset the main food groups are numbered from 1 to 61, but there is no food group associated with 46 [182]. This is due to changes in the coding of food groups made since previous versions of the survey.

Further insight into the results of a topic model can be gained by considering how the probability of each eating event type β_k varies with the time of day. As each collection of eating events is related to a time of day, for the purpose of visualisation, a 24 hour period can be split into 10 minute intervals and the collections of eating events grouped accordingly. The most probable eating event type across all collections of eating events for each 10 minute interval can be determined and plotted against the time of day. Furthermore, a more detailed graph can be generated, such as in figure 6.8, where the average probability of each eating event type across all documents in a 10 minute interval is plotted for every interval.

A quantitative assessment using a ten-fold cross validation was conducted, with perplexity (equation 2.15) as a measure of the performance of each model on the held-out fold. A fold is 10% of the data, with 103 randomly selected collections of eating events in each fold. For every fold a topic model was estimated using the remaining 9 folds of the data for 5 to 50 eating events types at sampling intervals of 5. The model was then used to perform inference on the held-out fold of data and the perplexity of the results was calculated. The average perplexity across all folds was then calculated for the range of number of eating event types. The number of eating event types around which the perplexity was lowest was identified and a more detailed analysis was carried out using the same method but for a range of eating event type numbers from 5 to 20 at sampling intervals of one. This was to provide a clearer picture of how the perplexity varies around this key value. This quantitative analysis ensures that the results are representative of the data and helps to confirm the selection of the number of eating event types determined by the qualitative analysis.

6.2.1.2 Sensitivity analysis for NDNS RP dataset

Whilst exploring the results of the topic models for qualitative analysis it was noted that certain vocabulary items often appear with a larger probability than others. In particular, the food group of ‘tea, coffee and water’ often dominates several topics in a model. This was highlighted by a nutrition expert to possibly be an artefact of the original dataset. Two different potential causes were considered and investigated further to understand their impact, if any, on the results of the model.

One possible reason for an elevation in probability of certain food groups is the coding method used in the NDNS RP dataset e.g. the nature of the groupings of food items into food groups. Each food group contains a different number of food items. For example, the ‘vegetables not raw’ food group includes many different vegetables, whereas the ‘eggs and egg dishes’ food group contains a smaller variety of items, such as omelette or quiche. Hence, if a person ate a roast dinner for example, the food group ‘vegetables not raw’ would occur 3 times in the relevant document if they had broccoli, cabbage and carrots as part of their meal. Whereas, if a person ate a quiche for example, the food group ‘eggs and egg dishes’ would only occur once in the document. This is not necessarily the most representative approach for nutrition analysis. To investigate the impact of this on the results a modified dataset of collections of eating events was created as before, but for every eating event each food group was only allowed to occur a maximum of once. Any repetitions of food groups in a single eating event were removed, although the collection of eating events could still have each food group listed multiple times overall as it still contains multiple eating events.

A second potential reason is the limited selection of food groups related to drinks appearing in topics compared with possible permutations of food groups related to food items. Hence, the same drinks will often occur with many different permutations of food. The impact of this potential bias was investigated by removing all eating events that contain only drinks, including sugar in teas and coffees and powdered drinks, such as hot chocolate. All dietary supplements were also removed from the dataset, as these are often taken as part of an otherwise drink only event and are not of particular interest in this analysis. A modified dataset of collections of eating events was created as before using this reduced set of eating events.

Finally, a third modified dataset was generated using collections of eating events where both the food groups could only be listed once per single eating event and where all drink only eating events were removed. Topic models were estimated for each of the three modified datasets using 10 topics and an initial α of 0.1, as these parameters

Table 6.2: Mappings of topic numbers to four eating event type categories for four modified datasets

Discovered eating event type no.	**Mapped eating event type label**			
	All data	No repeat	No drinks	No drinks or repeat
1	Main meal	Light meal	Light meal	Breakfast
2	Breakfast	Light meal	Main meal	Breakfast
3	Light meal	Breakfast	Light meal	Breakfast
4	Main meal	Snack	Main meal	Light meal
5	Snack	Snack	Breakfast	Snack
6	Light meal	Breakfast	Main meal	Light meal
7	Snack	Snack	Breakfast	Light meal
8	Snack	Main meal	Snack	Main meal
9	Breakfast	Main meal	Snack	Main meal
10	Main meal	Main meal	Main meal	Main meal

gave the lowest perplexity for the original full dataset and therefore allows comparison with the results from each modified dataset. The discovered eating event types for each dataset were visualised and evaluated qualitatively and were all considered to be representative of typical eating events. To compare the results, each of the 10 discovered eating event types in the four datasets were subjectively mapped by the author, based on their content, to one of four eating event type labels: 'Main meal', 'Light meal', 'Breakfast' and 'Snack' as shown in table 6.2

For each of the four eating event type labels the top food groups for all of the discovered eating event types mapped to the label were used to compare the probability distribution over these most relevant food groups. These distributions were compared visually by plotting them for each relevant eating event type label on one graph, as shown in section 6.2.2.2. Furthermore, a ten-fold cross validation was conducted for each dataset. The results were used to calculate the average perplexity across all folds for a range of numbers of eating event types and were compared across datasets.

6.2.2 Analysis and discussion of detected eating event types

6.2.2.1 Qualitative and quantitative results for full NDNS RP dataset

The semantic quality of topics in the context of investigating different types of eating events, was an important criteria when selecting the document structure and vocabulary. The results from a model created using the chosen document structure and

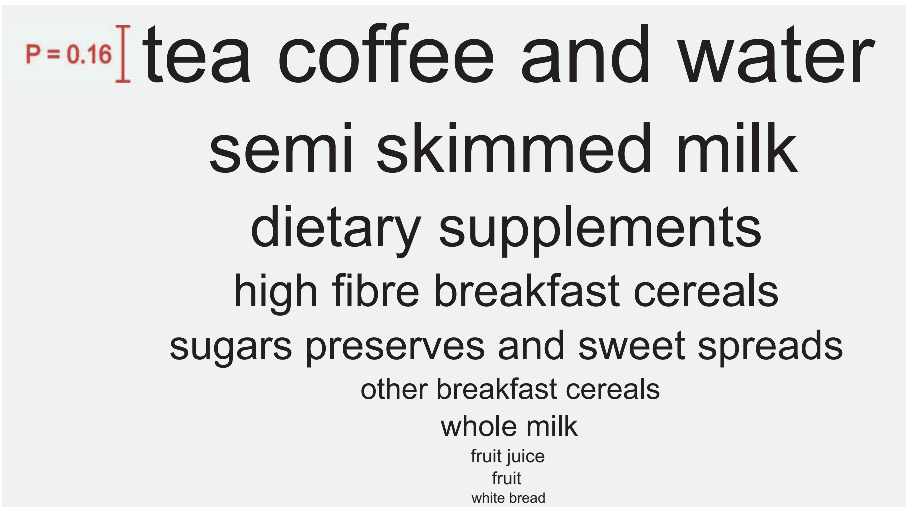


Figure 6.5: Top 10 food groups from an example 'breakfast' eating event type from a model with 15 eating event types, $\alpha = 0.01$. Displayed with font size proportional to probability of food group in the eating event type. Probability of first food group is given for scale.

Figure 6.6: Top 10 food groups from an example 'snack' eating event type from a model with 15 eating event types, $\alpha = 0.01$. Displayed with font size proportional to probability of food group in eating event type. Probability of first food group is given for scale.

Figure 6.7: Top 10 food groups from an example 'mixed' eating event type from a model with 15 eating event types, $\alpha = 0.01$. Displayed with font size proportional to probability of food group in eating event type. Probability of first food group is given for scale.

vocabulary, with 15 eating event types and an initial α set to 0.01, were considered to have a high semantic quality. Examples of eating event types detected by this model are given in figures 6.5 - 6.7. The full results are provided in appendix C.2. The discovered eating event types can vary widely in their semantic quality; some are very obviously representative of a particular type of eating event, whereas others are more ambiguous. The three examples shown have been assigned a subjective label to best describe their content. The eating event type in figure 6.5 is representative of breakfast and has a high semantic quality.

The eating event type in figure 6.6 is also semantically coherent and appears to reflect a snacking event. Finally, the eating event type in figure 6.7 appears to be of a lower semantic quality as it involves a wide mixture of food groups, which have relatively low probabilities, as shown by the size of the fonts. However, such discovered combinations of food groups can also provide useful insight as they may reveal patterns which are common but not necessarily well known or intuitive.

Figures 6.8 and 6.9, show how the fifteen eating event types found by this model vary over time, demonstrating the probability of eating event types at different times of day. It should be noted that the labels for the eating event types were assigned subjectively based on their most probable food groups. Furthermore, the frequency of eating events is not considered, hence a high probability of an eating event type at 2am does not necessarily correspond to a large number of this type of eating event. Rather, of those eating events that do occur, it is most likely to be of this type. For some times of day no eating events occurred and therefore a probability cannot be given.

It can be seen in figure 6.8 that during the early morning period the eating event types labelled as 'breakfast' have a high probability in comparison to other topics. These breakfast eating event types then generally have a lower probability at all other times. Similarly, the eating event types labelled as 'lunch/sandwiches' have peaks in probability around the middle of the day. However, these peaks are not as defined as those for breakfast. This suggests that there is more uncertainty in the type of eating events that occur at this time of day. In fact, the eating event types labelled

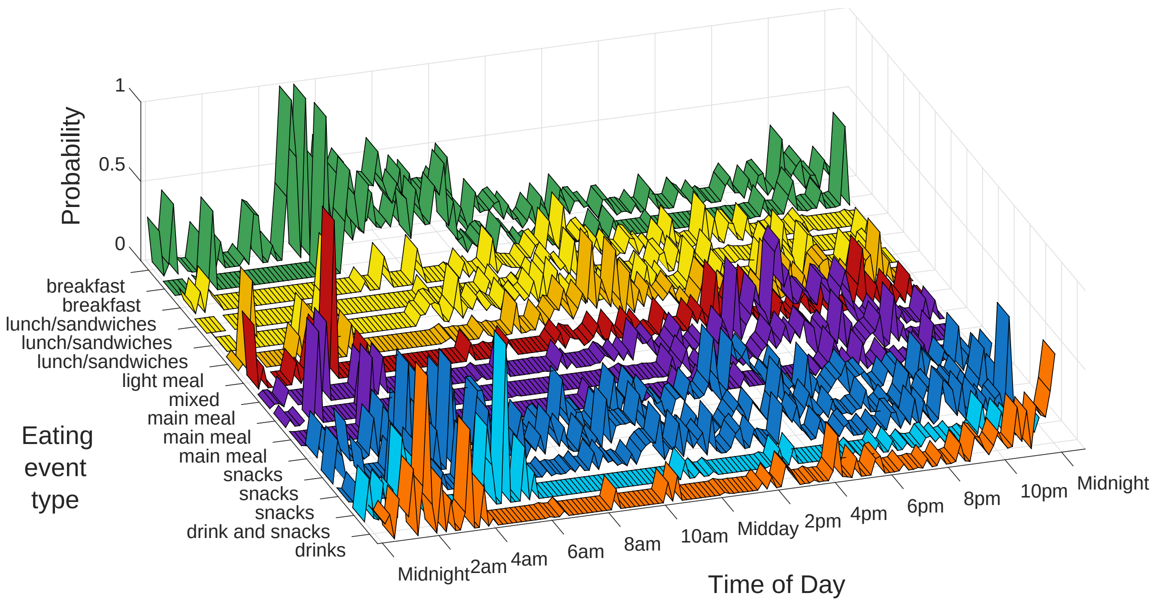


Figure 6.8: Graph showing variation of probability of each eating event type with time of day for a model with 15 eating event types, using the selected document structure and vocabulary for full dataset.

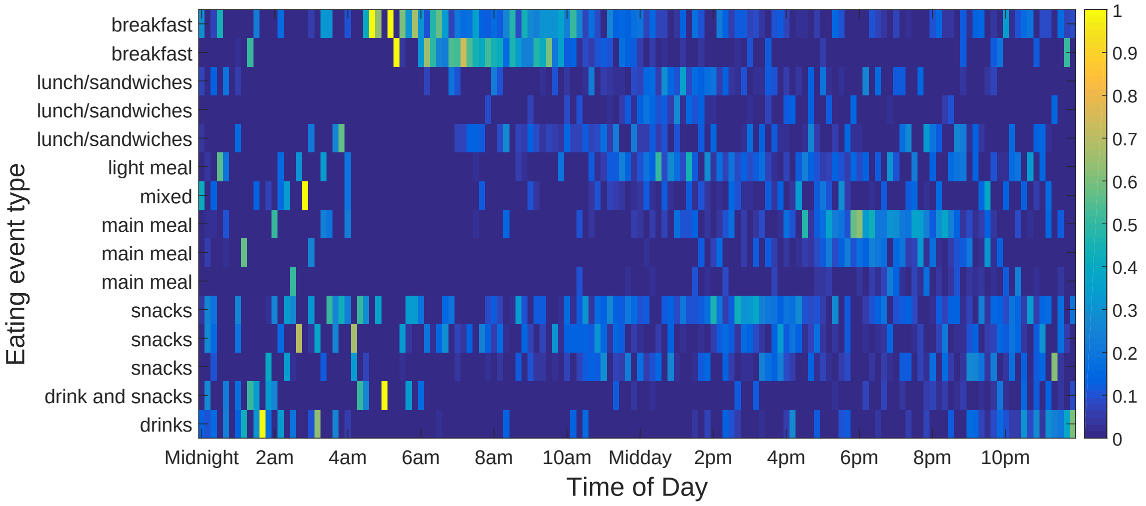


Figure 6.9: Matrix representation of how probability of each eating event type varies with time of day for a model with 15 eating event types, using the selected document structure and vocabulary. The colour indicates the probability.

as ‘snacks’ are also quite probable at this time. This is probably a combination of people consuming what is considered a snack rather than lunch and the wide range of times over which lunch type eating events occur. Overall, the snacks topics are more variable in probability throughout the full 24 hour period, this seems to fit with the cultural idea of snacking.

Figure 6.9 highlights the peaks in probabilities more clearly, by displaying a matrix representation of the eating event types against time, where the colour corresponds to the probability, as given by the values in the legend. For example, it can be seen that the two types of eating events labelled as ‘breakfast’ have distinctive peaks in probability at different times. One has a dominant peak around 7 - 8am, whereas the other is more dominant either side of these times. This difference could highlight a

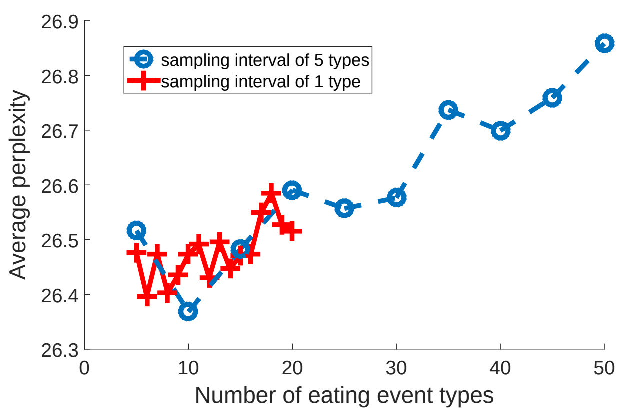


Figure 6.10: Quantitative results showing the variation of the average perplexity with the number of eating event types at different sampling frequencies.

link between the type of breakfast and the time consumed. It can also be seen that one of the main meal type of eating events has stronger probabilities than the others. These visualisations and observations provide a starting point for further analysis by nutritionists.

Figure 6.10 displays the results of the ten-fold cross validation conducted for the selected document structure and vocabulary at sampling intervals of 5 eating event types. It can be seen that at this resolution the average perplexity is lowest for 10 eating event types. The more detailed analysis based on this value, with sampling intervals of 1 eating event type, shows that the average perplexity oscillates between 5 and 16 eating event types. However, it should be noted that the scale of the average perplexity has a small range (0.6) and therefore the number of eating event types does not have a strong effect on how well the model generalises to previously unseen data. The average perplexity for 15 eating event types, the number chosen based on the qualitative analysis, is very similar to the lowest average perplexity indicating that using 15 types is valid.

For completeness, as the minimal average perplexity was for a model with 10

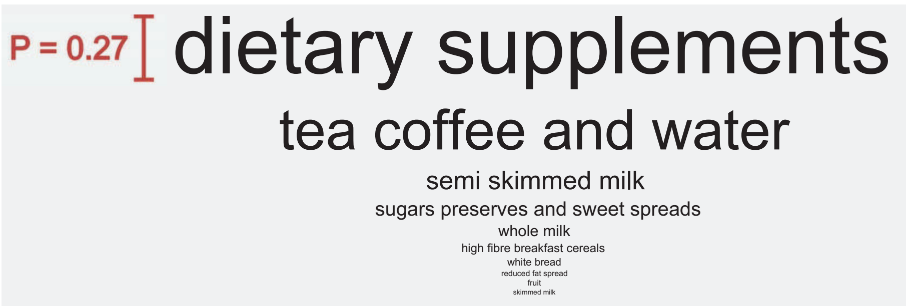


Figure 6.11: Top 10 food groups from an example 'breakfast' eating event type for a model with 10 eating event types. Displayed with font size proportional to probability of food group in eating event type. Probability of first food group is given for scale.

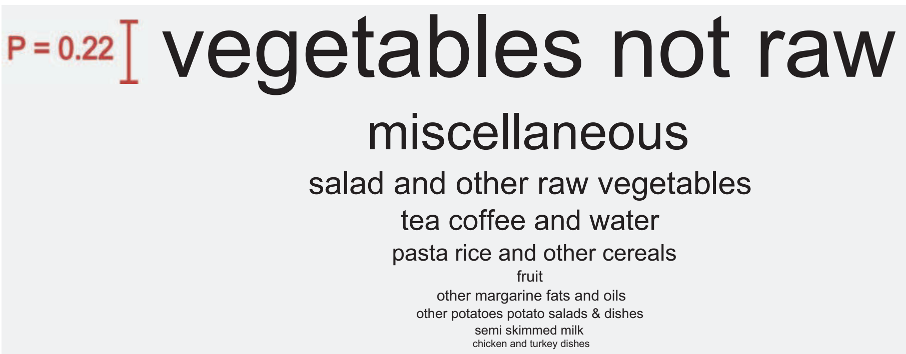


Figure 6.12: Top 10 food groups from an example 'main meal' eating event type for a model with 10 eating event types. Displayed with font size proportional to probability of food group in eating event type. Probability of first food group is given for scale.

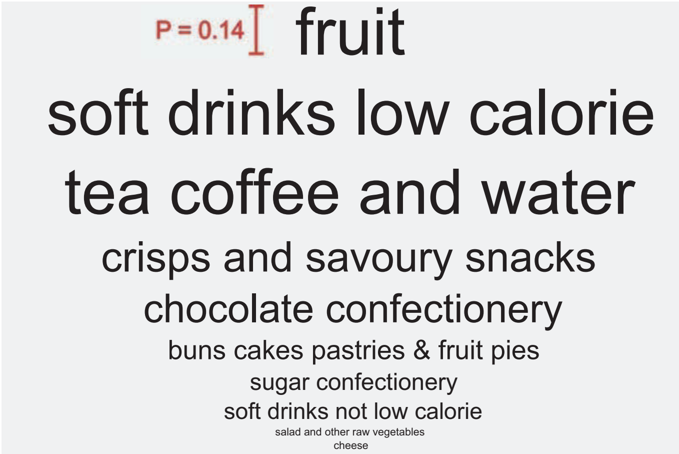


Figure 6.13: Top 10 food groups from an example 'snacks' eating event type for a model with 10 eating event types. Displayed with font size proportional to probability of food group in eating event type. Probability of first food group is given for scale.

eating event types, examples from this model are given in figures 6.11 - 6.13. The full results are provided in appendix C.1. The top 10 food groups are displayed with font size proportional to the probability of the food group in the eating event type and based on these a subjective label is assigned. Figure 6.11 shows an eating event type labelled as 'breakfast', however it can be seen that this is quite a different type of breakfast, compared to figure 6.5 as it is dominated by food groups relating to a cup of tea or coffee and taking dietary supplements. Figure 6.12 gives an eating event type labelled as a 'main meal', as it includes cooked vegetables and pasta or rice. Both of these eating event types are considered to have quite good semantic quality as it is relatively easy to apply a subjective label. The eating event type in figure 6.13 is labelled as 'snacks' and is considered to have a high semantic quality as several food groups have relatively high probabilities and the top eight words are typical of what are often considered snacking foods or drinks [202].

Figures 6.14 and 6.15 show the variation in probability of the 10 eating event types over time. These are very similar to the patterns seen for 15 eating event types. The structure of each type is listed in full for both models in appendices C.1 and C.2. Table 6.3 lists the ordered subjective labels used in figures 6.8 and 6.15 and the corresponding eating event type number for reference to the appendices. It can be seen in 6.14 that the breakfast eating event types have the highest probabilities between 6 and 10am. The probability of the snack eating event types are again more varied throughout the day and the main meal eating event types have peaks during the evening. Figure 6.15 highlights that different types of main meals appear to have stronger associations with different times of day, for example, an early dinner time around 5pm or a later meal between 7 and 8pm.

Table 6.3: Reference table for eating event type numbers for each model. Full listings of the content of each type can be found in Appendices C.1 and C.2.

Model with 15 eating event types		**Model with 10 eating event types**	
Ordered Label	Number	Ordered Label	Number
Breakfast	2	Breakfast	2
Breakfast	13	Breakfast	9
Lunch/sandwiches	5	Lunch/sandwiches	6
Lunch/sandwiches	11	Light meal	3
Lunch/sandwiches	15	Main meal	1
Light meal	8	Main meal	4
Mixed	4	Main meal	10
Main meal	6	Snacks	5
Main meal	12	Snacks	7
Main meal	14	Drinks/snacks	8
Snacks	1		
Snacks	7		
Snacks	10		
Drinks/snacks	9		
Drinks	3		

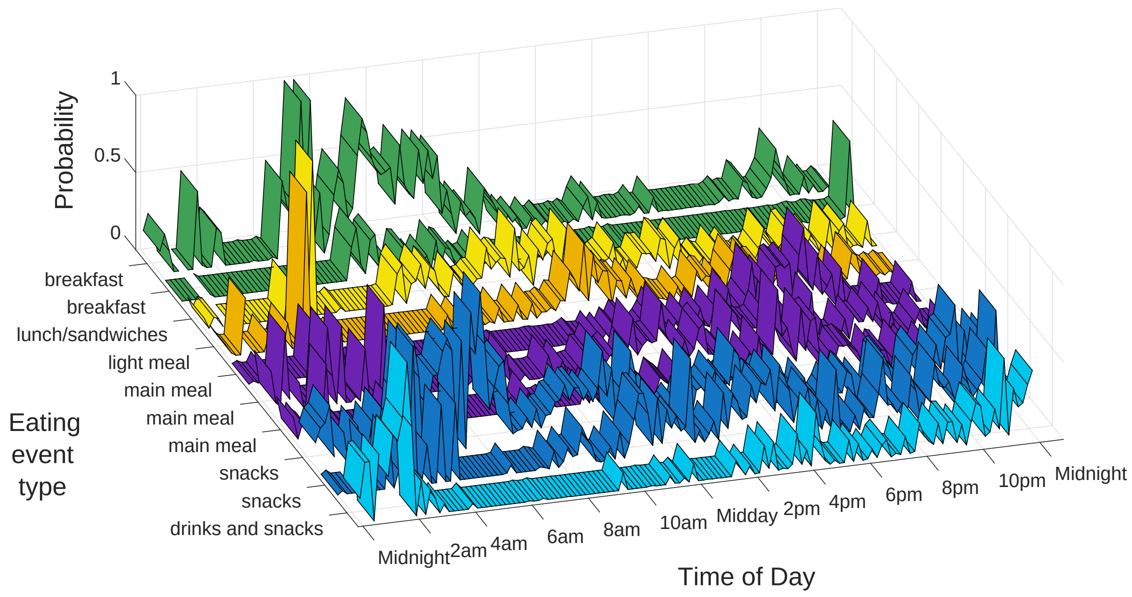


Figure 6.14: Graph of variation in probability for eating event types with time of day for a model with 10 eating event types, using the selected document structure and vocabulary.

6.2.2.2 Sensitivity analysis results

Figures 6.16 - 6.19 show the comparison of eating event types mapped to each of the labels for all of the datasets: original (all data), food group repetitions in single eating events removed (no repeat), drink only events removed (no drink) and both

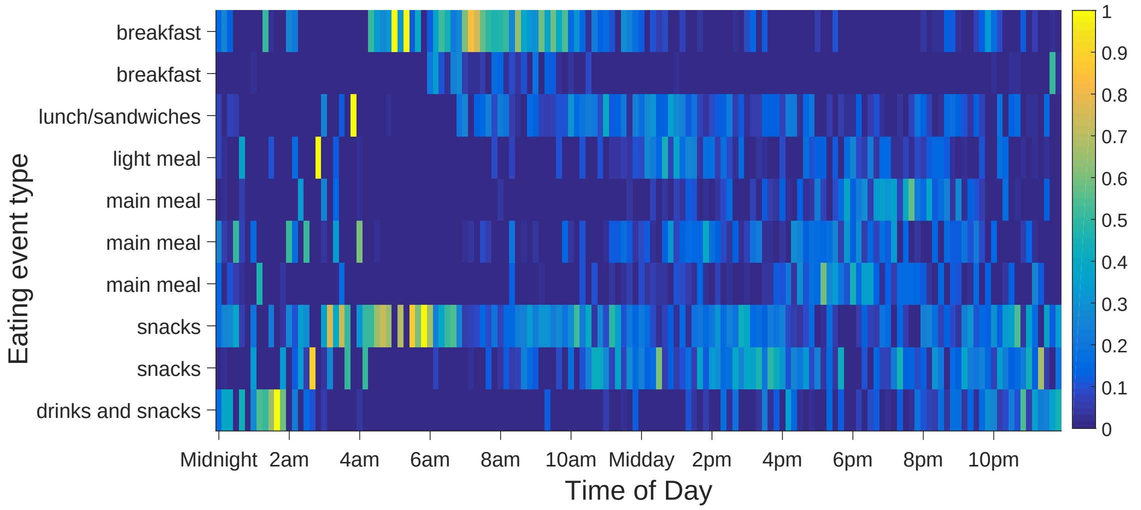


Figure 6.15: Matrix representation of how probability of each eating event type varies with time of day for a model with 10 eating event types, using the selected document structure and vocabulary. The colour indicates the probability.

repetitions and drinks removed (no drinks or repeat). For each dataset 10 eating event types are detected and subjectively mapped to the four labels. The number of discovered eating event types that match to a specific label will vary between datasets. Each label can have more than one eating event type mapped to it for each dataset, therefore these are numbered successively. For example, the dataset with drink only events removed has three eating event types mapped to the 'Breakfast' label, hence these are referred to as 'No drinks 1', 'No drinks 2' and 'No drinks 3'.

Figure 6.16 shows that for the 'breakfast' label, the distributions over the vocabulary appear to vary as much between two eating event types created from the same dataset as they do between eating event types from different datasets. In particular, looking at the 'tea, coffee and water' food group, it can be seen that this has a strong probability in several eating event types, including those from datasets which had drink only events removed. All of the datasets contain at least one eating event type with a relatively high probability of food groups related to breakfast cereals and another with a low probability for the same food groups. Similarly for the 'light meal' label, figure 6.17 shows that the variation in distributions is not clearly related to the dataset. Again, removing drink only events does not necessarily reduce the probability of the 'tea, coffee and water' food group. Removing repetitions of food groups within an eating event may have had an impact on the probability of the food group related to salad, however this can also have a low probability in topics from datasets where this restriction was not applied.

The range of distributions over the vocabulary does not appear to be linked to the dataset for the 'main meal' label either, as demonstrated by figure 6.18. For

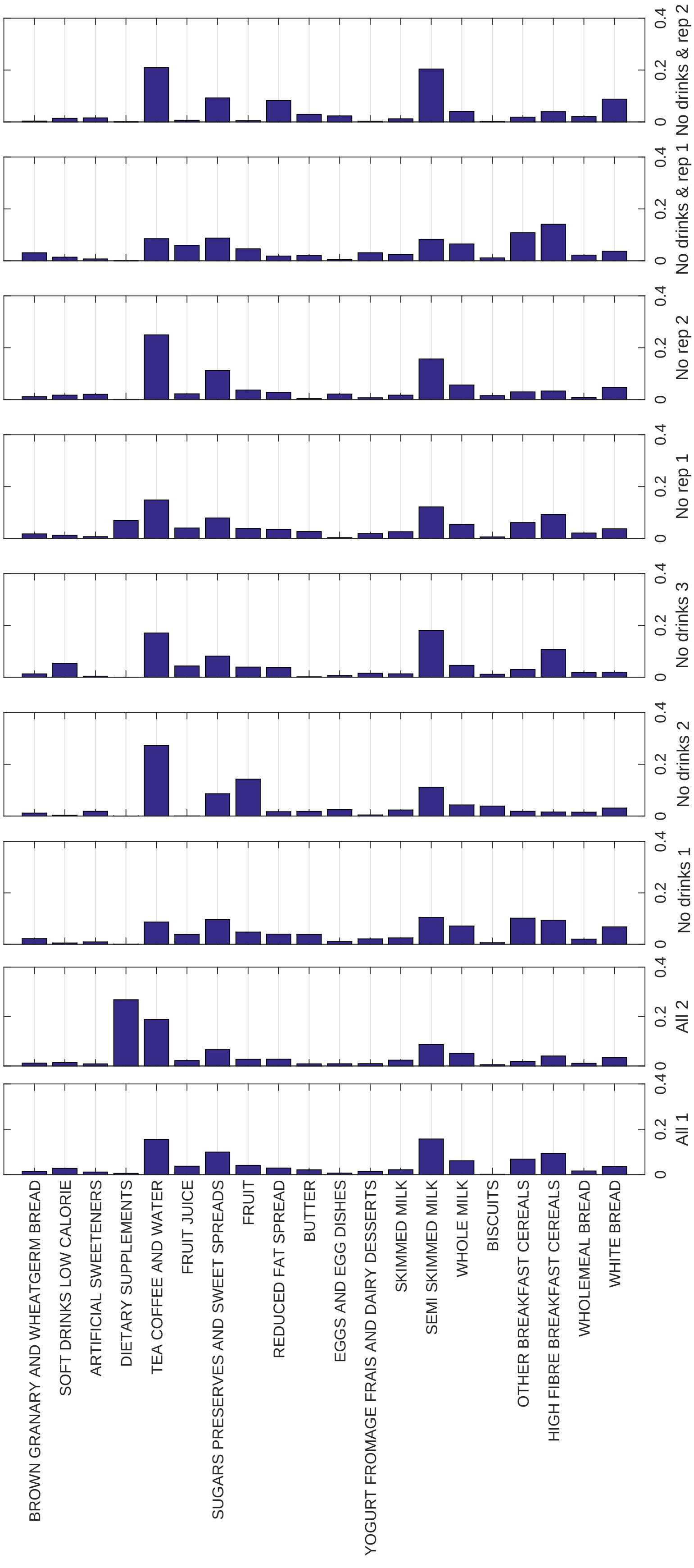


Figure 6.16: Comparison of eating event type distributions from four different datasets over relevant vocabulary items for the breakfast eating type event label.

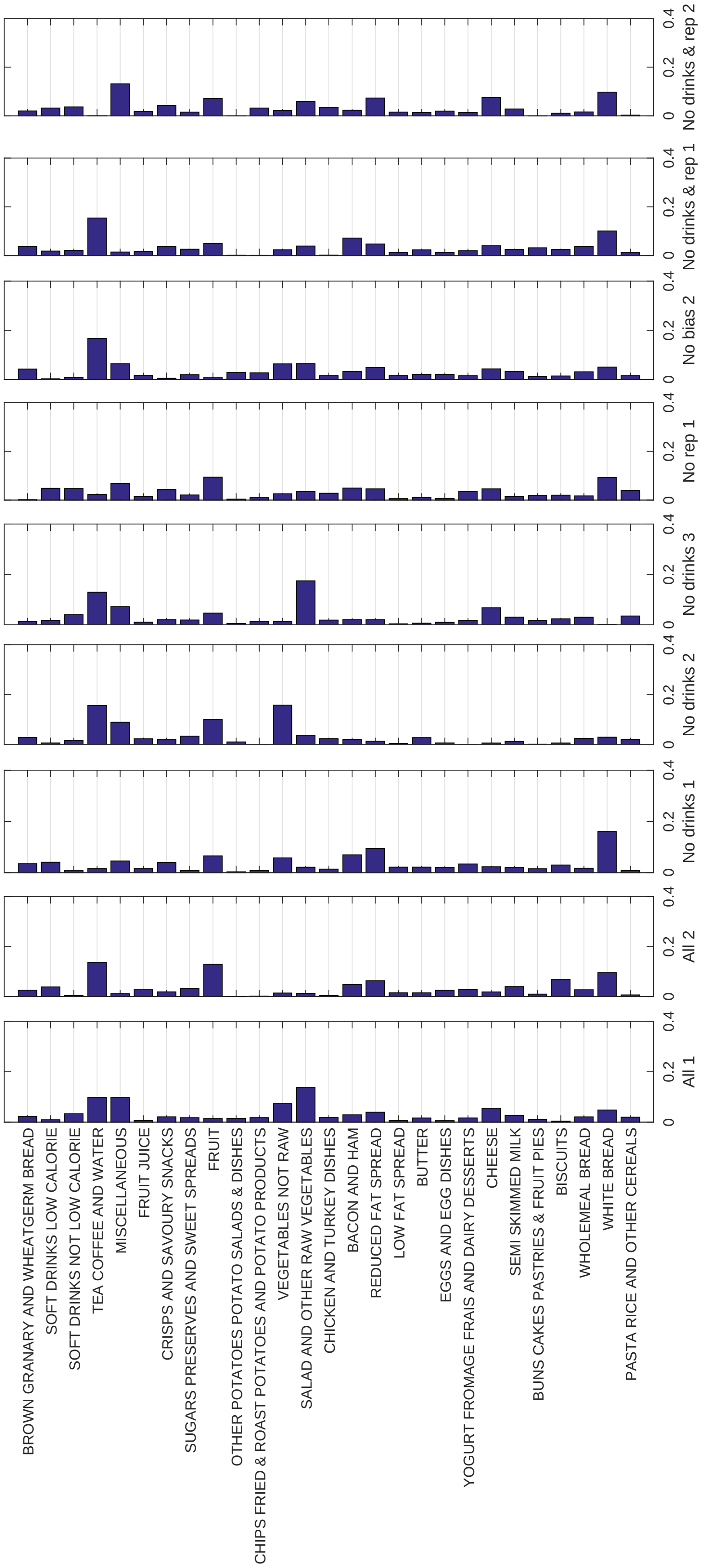


Figure 6.17: Comparison of eating event type distributions from four different datasets over relevant vocabulary items for the light meal eating type event label.

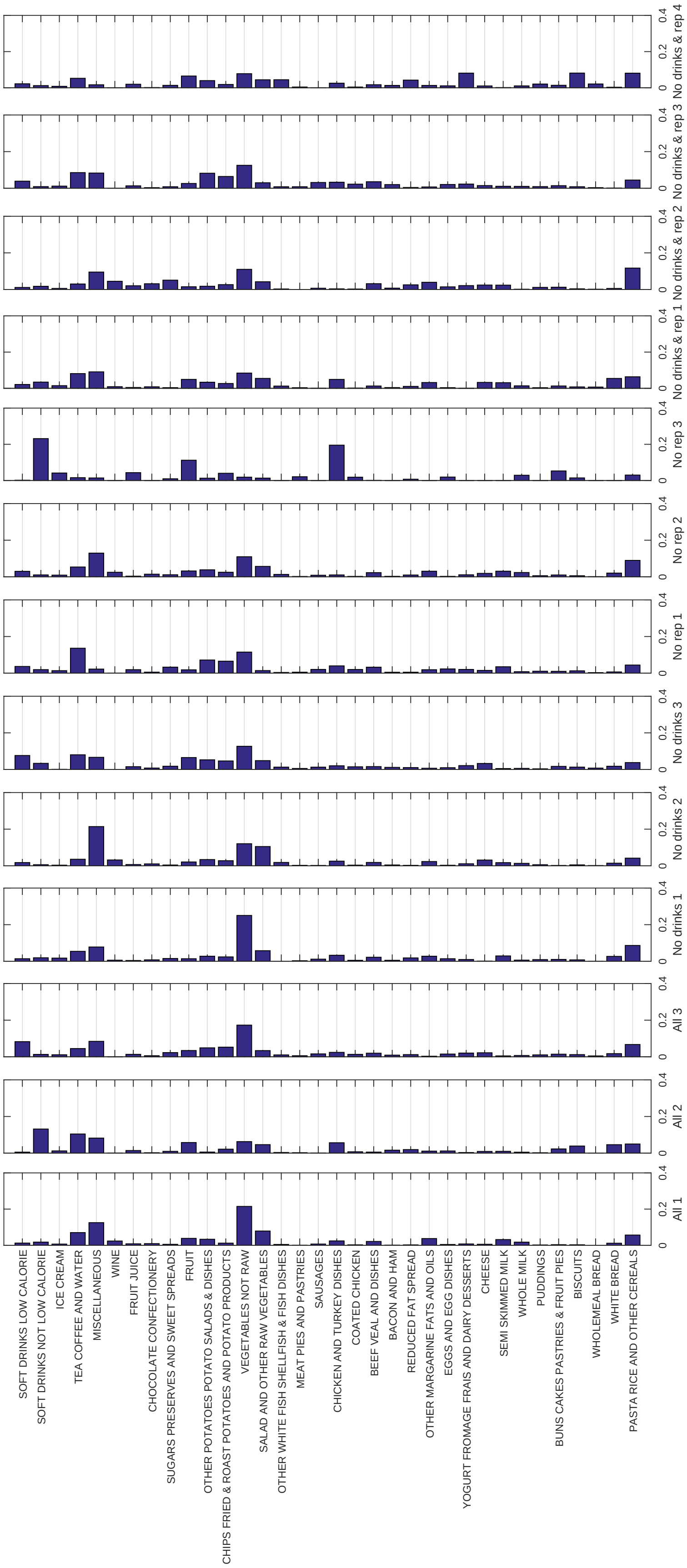


Figure 6.18: Comparison of eating event type distributions from four different datasets over relevant vocabulary items for the main meal eating type event label.

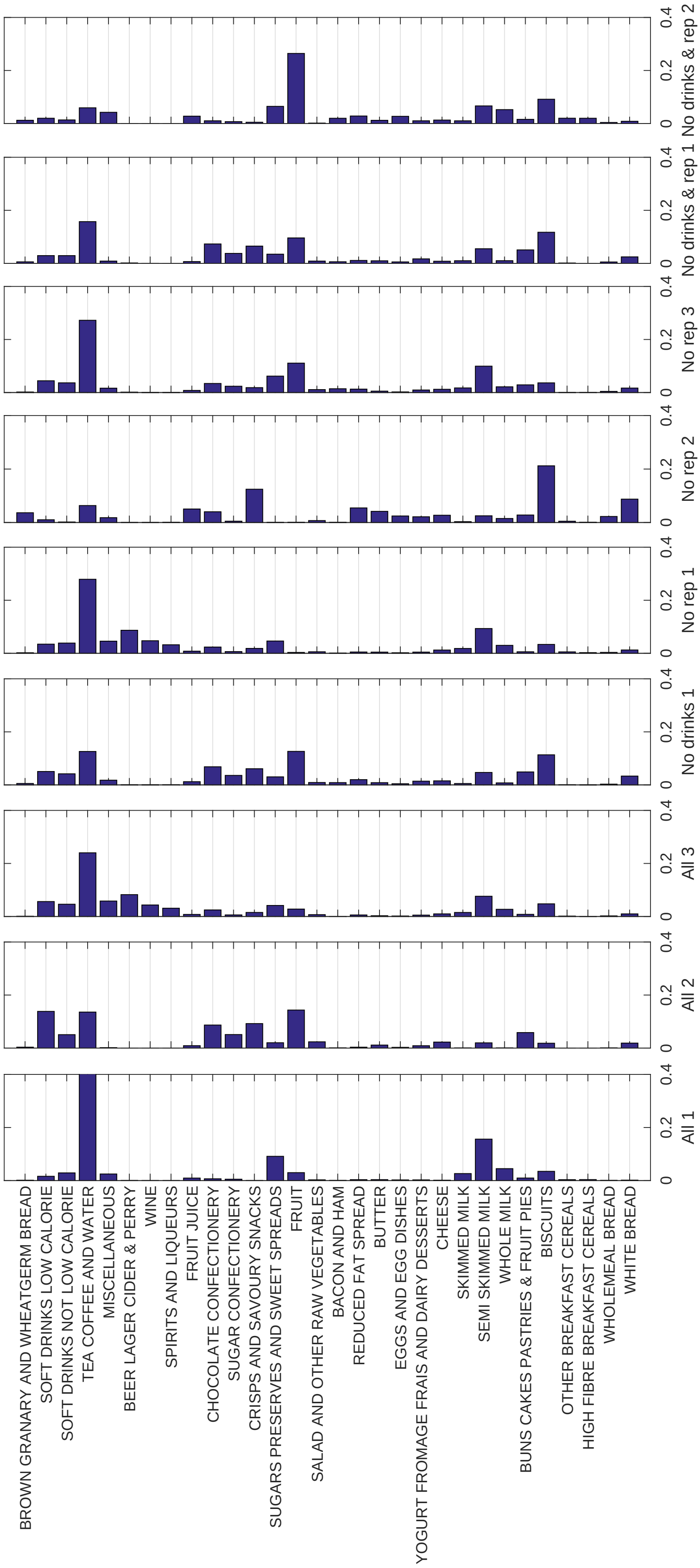


Figure 6.19: Comparison of eating event type distributions from four different datasets over relevant vocabulary items for the snack eating type event label.

example, there are several eating event types with a focus on food groups related to potatoes and vegetables from a variety of datasets. There are also eating event types with a focus on pasta/rice and vegetable food groups in each dataset, although the probability of vegetables is a little lower for the eating event types from datasets with repetitions removed. Figure 6.19 highlights that there are eating event types with the 'snacks' label with similar distributions estimated from different datasets. For example, the third eating event type from the dataset using all the original data (All 3) and the first eating event type from the dataset with only repetitions removed (No repeat 1) are very alike.

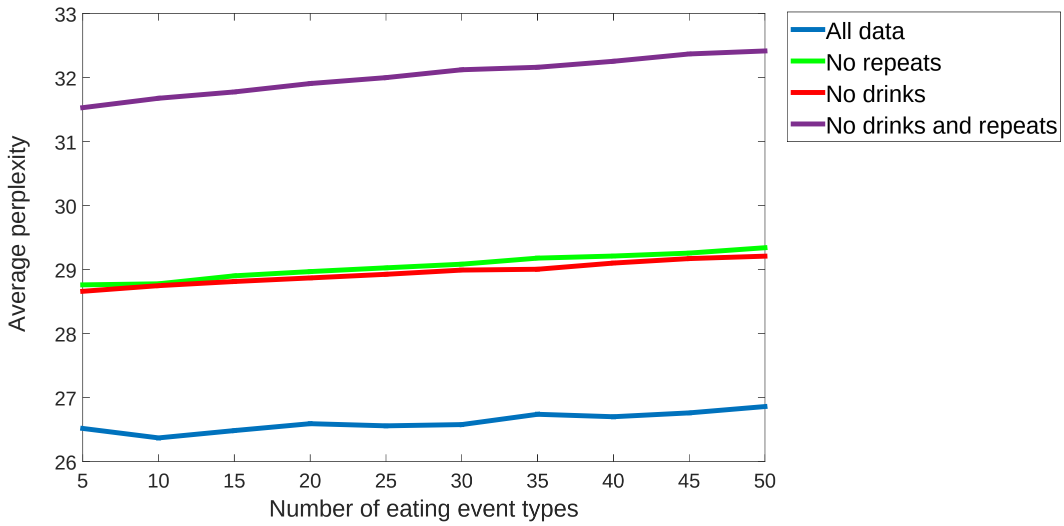


Figure 6.20: Graph showing how the average perplexity varies with the number of topics for four modified datasets

As well as the qualitative comparison of the different datasets, figure 6.20 shows the results of the ten-fold cross validation performed on each dataset. It can be seen that removing either repeated food groups or drink only events corresponds to an increase in the perplexity across the range of number of eating event types. Furthermore, removing both repetitions and drink only events increases the perplexity further again. Hence, using all of the original data gives the best generalised performance on previously unseen documents.

6.3 Chapter summary

This chapter has considered the application of LDA to a different data source, namely nutrition data, with the aim of revealing patterns in the combinations of food groups consumed together in different types of eating events. The critical importance of the selection of a vocabulary and document structure for the LDA model when using a new type of data has been shown. In particular, the exact context and aim of the task should always be taken into account. One key result that can be applied to other data types is to ensure that all of the assumptions of LDA are completely met, in particular, that documents exhibit multiple topics. This will depend on what a topic has been defined to be for a particular application but should be given careful consideration. Moreover, the selection of the vocabulary will have a strong impact on the semantic quality of the resulting topics. For example, a large vocabulary may not be suitable as the resulting topics are likely to not be specific enough. It is important to consider the choice of vocabulary in the context of the application and to recognise that documents use a bag of word representation which has no structure modelled.

It has been demonstrated by the results presented in this chapter that it is possible to apply topic models to nutrition data from food diaries in order to find topics that are representative of different types of eating events. This was done using the NDNS RP dataset, which contains a sample of food diaries representative of the UK population. Different methods of visualising the resulting eating event types have been shown. This includes listing the top food groups of an eating event type with the font size in proportion to the probability of the food group and displaying the variation in probability of eating event types with the time of day. This has confirmed that the topics found intuitively correspond to typical eating event types, as confirmed by a nutritionist expert with experience in looking at different combinations of food groups in the context of eating events.

Finally, a sensitivity analysis of the effect of different properties of the original dataset was conducted. In particular, repeated food groups in a single eating event and drink only events were removed from the data and the results were compared. Overall, from both the qualitative and quantitative analysis there is no evidence that altering the dataset properties improves the performance of the topic models. Therefore, as making changes could cause unforeseen consequences it was concluded that utilising the full dataset is the most appropriate option.

Chapter 7

Exploring and Validating Eating Event Types

In chapter 6, it was demonstrated that topic models can be applied to nutrition data to find topics representative of types of eating events. This initial research was done using the NDNS RP dataset. This chapter extends this work by considering further visualisations and analysis that can be conducted to exploit these results. For example, the dataset can be separated into subsets based on the type of day, weekday or weekend day, in order to compare the number of each type of eating event within these groups. In addition, inference can be performed on a per person basis to look at the specific eating event types for an individual. The aim is to highlight a variety of different ways that the eating event types discovered by topic models can be used for nutrition research, rather than answering a specific nutrition based research question.

Furthermore, the application of topic models to two other existing nutrition datasets is considered to validate that the methods for selecting a vocabulary and document structure extend to datasets with different properties. Alternative approaches to identifying eating event types have previously been used with the selected datasets and can be compared with the results from topic models. Firstly, a manual rule-based labelling approach to identify types of eating event was previously applied to the UK National Diet and Nutrition Survey (NDNS) 2000 dataset [202]. Secondly, topic models are implemented for the Irish National Adult Nutrition Survey (NANS) [183] because this dataset includes subjective labels for each eating event recorded as part of the original data collection procedure.

7.1 Exploring the NDNS RP dataset using discovered eating event types

The eating event types that have been discovered using topic models can be used to analyse the NDNS RP dataset in more detail. The food group combinations found using the model with 15 eating event types, presented in section 6.2.2.1, are used as these were considered to be the most semantically coherent results. The eating event type labels assigned to the discovered combinations of food groups are also considered in the analysis, however these labels are used with caution due to their subjective nature.

7.1.1 Inference for individuals

One avenue for exploring the dataset using the results from the topic model is to perform inference at an individual person level. Documents can be created based on single eating events, although this does not match with the assumption that a document will exhibit multiple topics, these will only be used for inference, rather than estimating a model and allow an eating event type to be associated with every eating event. Each person will have a different number of documents, depending on the number of eating events they recorded across the four day food diary. For example, participant one recorded 27 eating events over a four day period.

7.1.1.1 Visualising the types of eating events for an individual

The results of inference at the individual level can be used to investigate an individual's eating behaviour patterns by visualising the type of eating events discovered. Figure 7.1 shows the most likely eating event type for every eating event recorded by participant one on the first day of their food diary. For every eating event the original food groups of the items consumed in that event, as listed in the food diary, the time of the event and the most likely eating event type found by inference, along with the corresponding subjective label are given. For example, at 13:30 the participant had a piece of fruit, a soft drink and two items of sugar confectionery. This was found to most likely be eating event type 10, which was subjectively labelled as a snack. Intuitively, this label seems to be a good match to the items consumed during the eating event.

The subjective labels for the eating events at 07:45 and 11:08 are less intuitive. The event at 07:45 is likely to be considered breakfast due to the food group com-

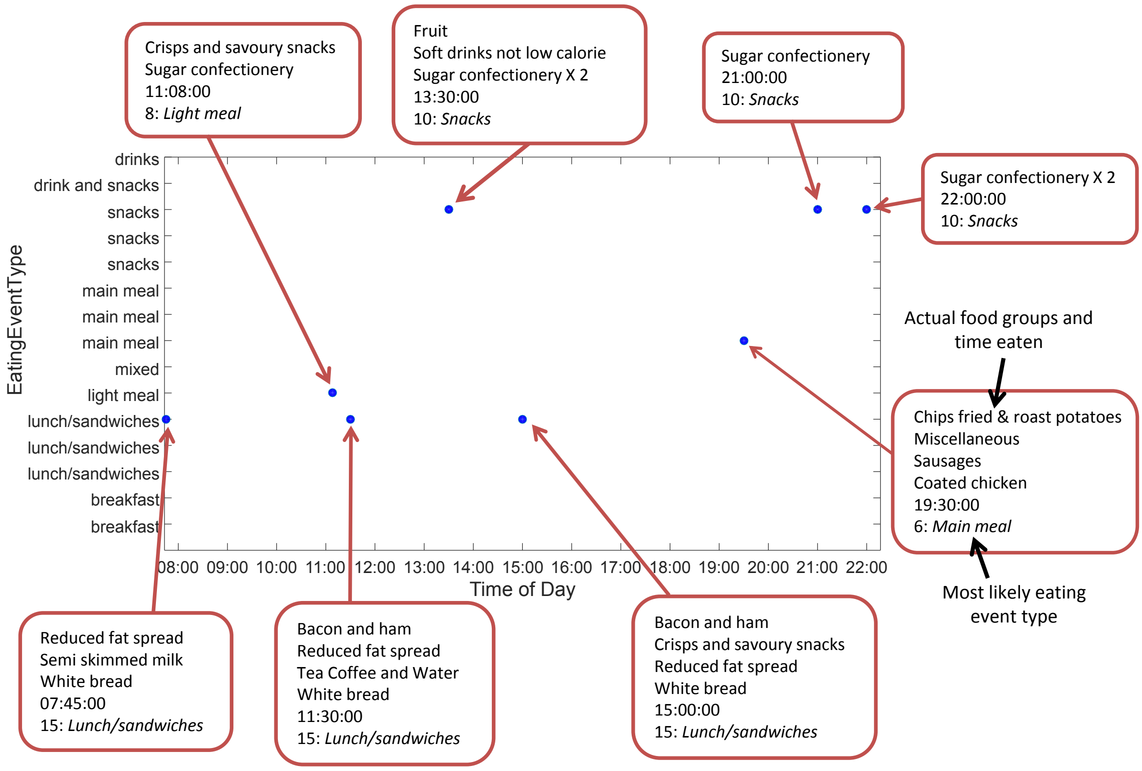


Figure 7.1: The most likely eating event type number and corresponding subjective label found by inference for each eating event recorded by participant one, on food diary day one. The recorded food groups and time of the eating event are given for context.

binations and the time of the event. However, it can be seen why the most likely eating event type found during inference was one that was subjectively labelled as lunch/sandwiches, because the bread and spread could easily be part of a sandwich rather than toast. Similarly, the event at 11:08 is assigned to most likely be eating event type 8, which is subjectively labelled as a light meal. This event has food groups that could form part of a light meal, however this is much more likely to be labelled as a snack by a participant or researcher.

The results for this participant highlight the complexity and subjective nature of analysing food and drink intake at a meal level due to different behaviour patterns. This participant has recorded three separate eating events between 11.30 and 15:00. Considering the content of these eating events it appears that the participant probably spread their 'lunch' over this extended period as a series of smaller eating events. In particular, bread, spread and bacon/ham are listed at both 11:30 and 15:00 suggesting that the participant ate one sandwich late morning and another in the afternoon. The total content of these three eating events could be equivalent to one eating event recorded by a different participant, either because they ate everything together or just recorded it as eaten at one time for ease. If the nutrition intake was considered at a day level this would be equivalent overall, however the behaviour of having a series of smaller eating events could be linked to health outcomes and hence this meal level information is important. The traditional labels of lunch and snack may not be appropriate for reflecting this eating behaviour pattern.

Figure 7.2 shows the eating events for day two of the food diary for participant one. This shows that, similarly to the first day, the subjective labels for the most likely eating event type correspond well to the content of each eating event. Again, one exception is the event that is most likely to be type 8 that occurs at 21:29, which also appears to be more likely to be considered a snack rather than a light meal. Comparing the results for the two days it can be seen that this participant has little consistency or routine in the type and time of their eating events across these two

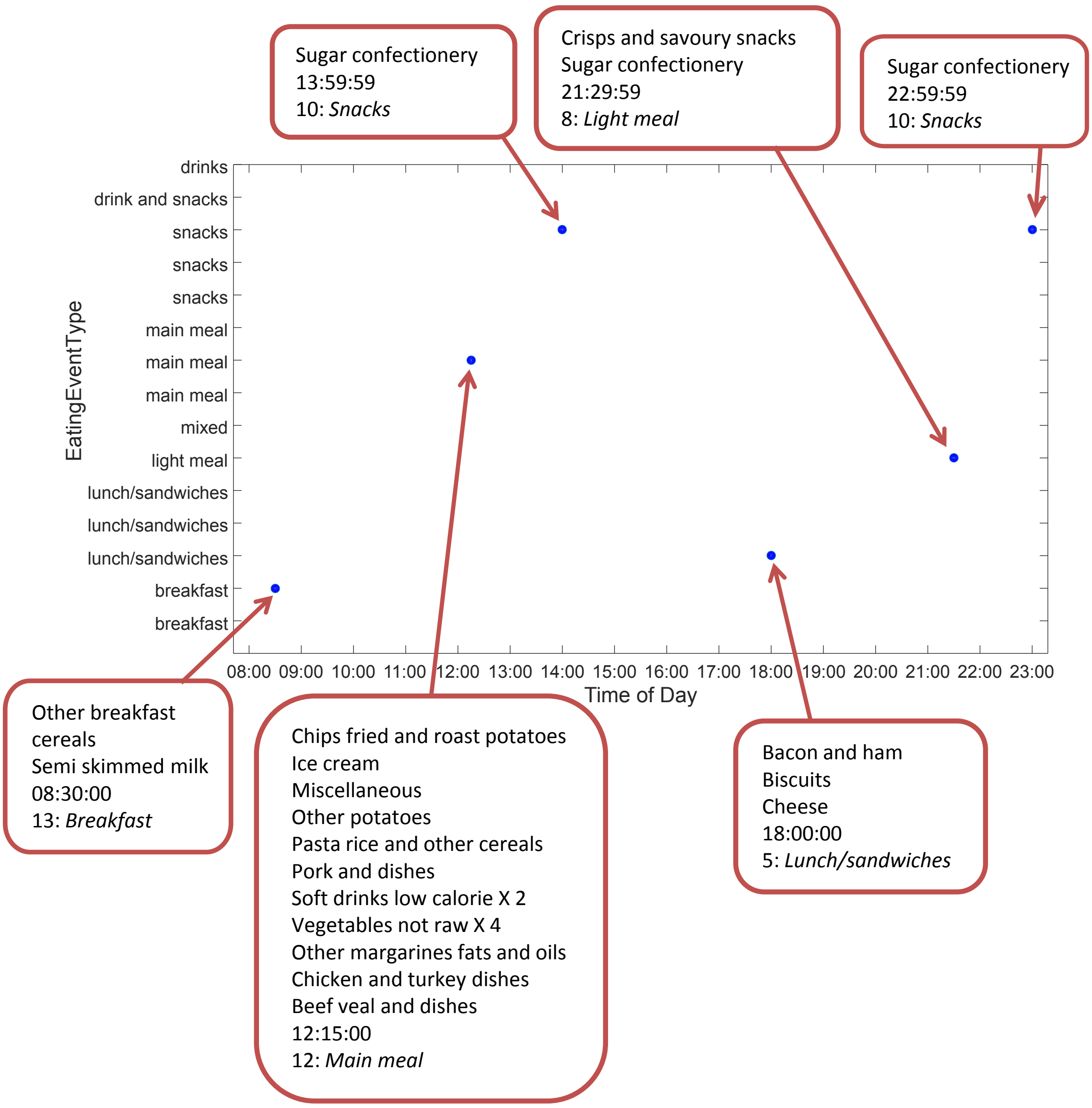


Figure 7.2: The most likely eating event type number and corresponding subjective label found by inference for each eating event recorded by participant on food diary day two. The recorded food groups and time of the eating event are given for context.

days. Referring back to the original data reveals that these entries are for a Saturday and Sunday respectively. This may influence the type and number of eating events that occurred [193].

Figure 7.3 highlights the most likely eating event types across the whole four days of participant one's food diary. It can be seen from this plot that there are no strong patterns of the time and type of eating events for this participant over the four day period. Figure 7.4 shows the eating events for a different participant for all four days of their food diary. It can be seen that this participant is very different from the first example. Firstly, this participant has far fewer eating events across the four days, a total of 14 compared to 27 for the other participant. Secondly, this participant demonstrates a much stronger routine and consistency in their eating events. Both the content and the time of their eating events are frequently very similar or the same. All of the subjective labels for the most likely eating event types for this participant match well to the items consumed during the events.

A random sample of the participants was selected and the detailed content of their eating events were investigated. For brevity visualisations of all of these are not included here as the aim is to demonstrate different types of analysis that the use of topic models can facilitate, rather than undertaking a full analysis of the dataset. It was found that overall the majority of the subjective labels for the most likely eating event types match well to the items consumed during the eating events. The

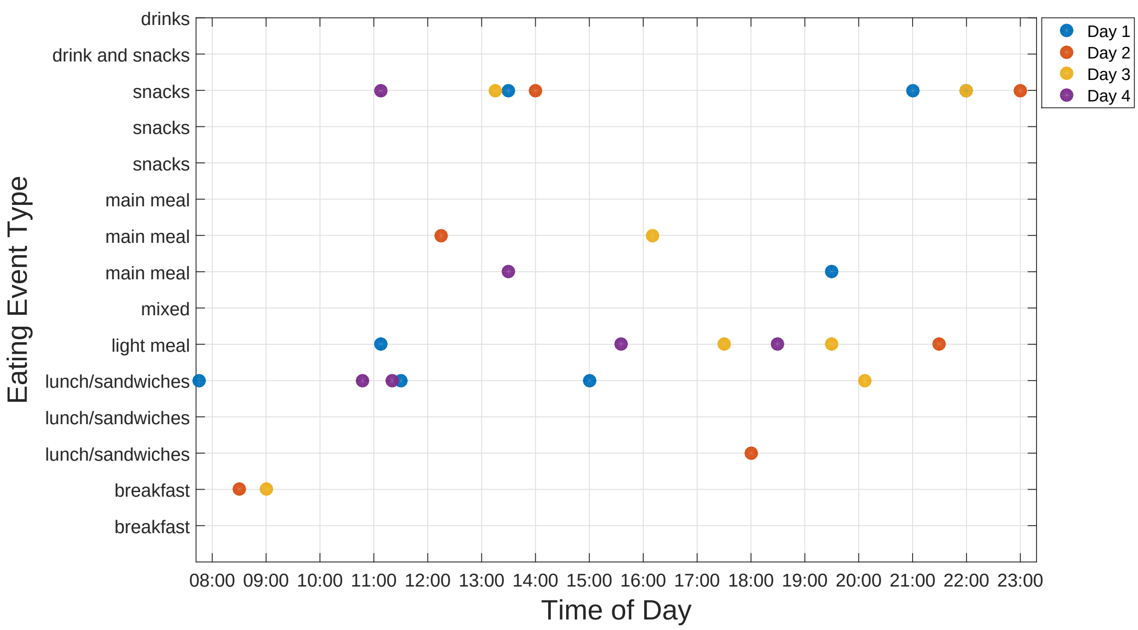


Figure 7.3: Graph showing times and most likely eating event type label found by inference for all eating events recorded by participant one, across all four days of their food diary.

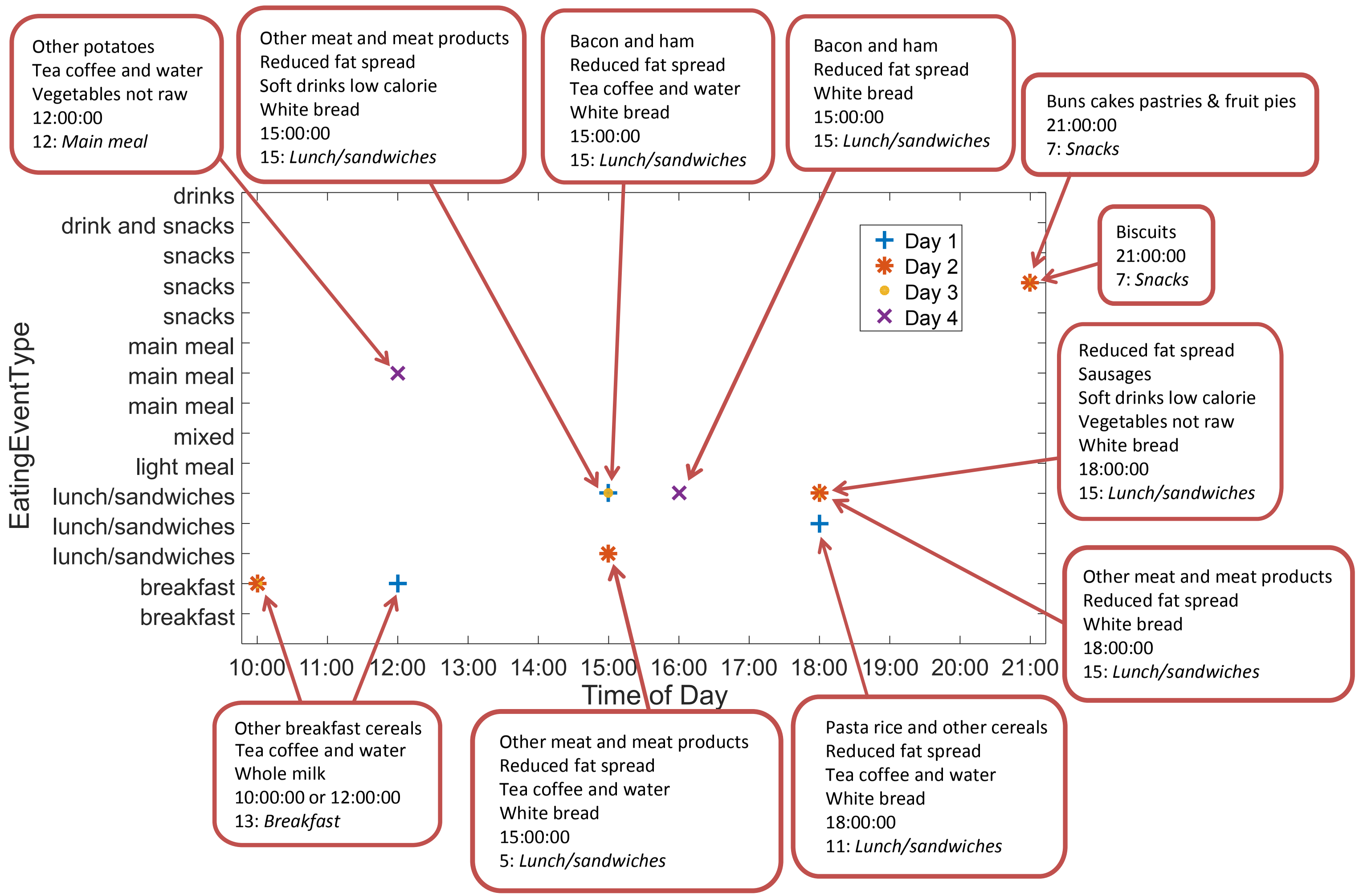


Figure 7.4: The most likely eating event type number and corresponding subjective label found by inference for each eating event recorded by a randomly selected participant across all four days of their food diary. The recorded food groups and time of each eating event are given for context.

exceptions found included an eating event containing only fruit, which was inferred to most likely be eating event type 15, lunch/sandwiches. Many people would probably consider one piece of fruit a snack, however it was consumed at 13:35 so the participant could have considered this lunch and hence the subjective label inferred would be correct. This highlights the lack of clarity of subjective labels. Another exception detected was the occurrence of eating events that contain the items used to bake a home made cake listed separately, which were inferred to most likely be a main meal. This is because the food groups of the recipe items do match the food groups likely to be in a main meal and the topic model has no other context. This problem could be solved by listing such items as the finished product, i.e. cake, rather than the individual ingredients.

7.1.1.2 Investigating the effect of morning eating event types

In addition to visualising the eating events for a specific individual, the results of inference at the individual level can be used for further analysis across the full sample population. For example, nutritionists are often interested in the effect of breakfast and how this influences what someone eats for the rest of the day [230]. To demonstrate how the results from topic models can provide a new approach to this research area, an initial analysis was conducted. No direct nutritional conclusions are drawn, as this is not the focus of this thesis, but this approach could be utilised by the nutrition research community to address relevant hypotheses.

The most likely eating event type for each eating event for all participants was determined using the individual inference results. For each day of every participant's food diary the eating events that occurred between 6am and 10am were selected and the most common eating event type amongst these was identified. Based on this information the eating events recorded for that day were put into one of three categories: none, breakfast or other. This reflected whether the participant had no eating events, mostly breakfast type events or mostly other types of events within this time period. The most likely eating event type for each of the remaining eating events for the rest of the day were then listed for the corresponding category.

Figure 7.5 shows the average number of each of the 15 eating event types per day for each of the three categories. Each of the categories: none, breakfast and other, contained 2997, 8401 and 5141 days respectively. It can be seen that the distribution of eating event types varies across the categories. Overall, the none category has lower average numbers of eating events across all types. This suggests that people who do not consume anything between 6am and 10am generally have fewer total number of eating events in a day than those who consume something in

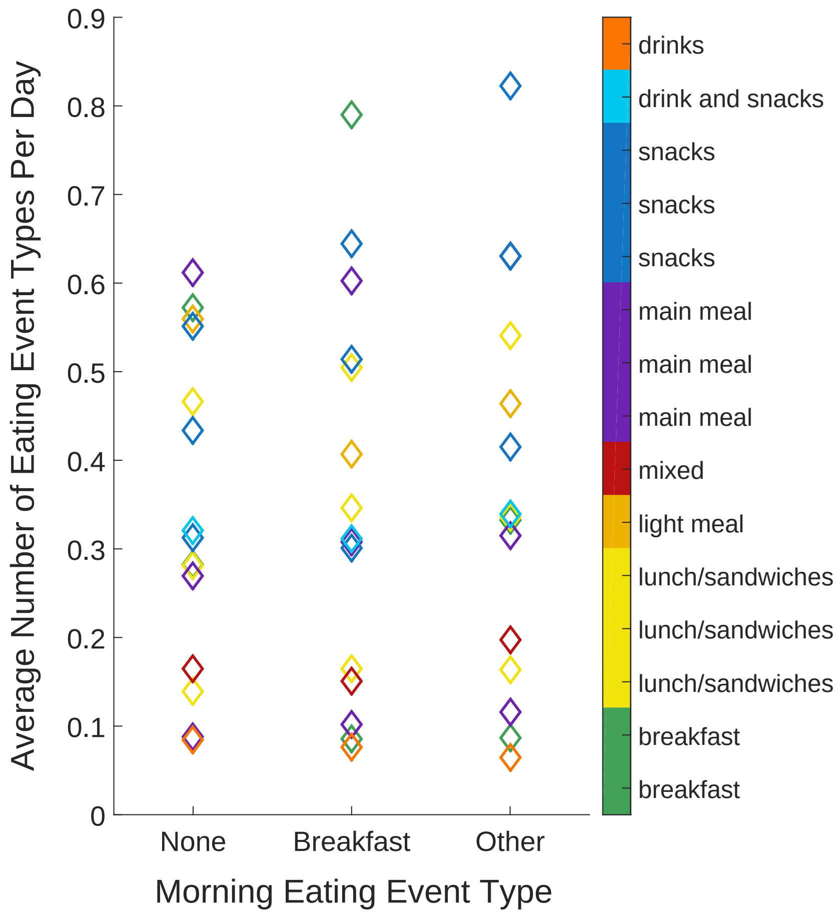


Figure 7.5: Graph showing the distribution of the average number of eating event types in one day for three categories based on the most common eating event types between 6am and 10am. All eating event types are found by inference at the individual level.

this time period. However, this analysis does not take energy content into account. Nutrition researchers could take this idea further to investigate whether fewer eating events correlate to a lower daily energy intake. Figure 7.5 also reveals that those who primarily have eating event types that are not breakfast between 6am and 10am have a higher average number of snack eating event types across the rest of the day. Again, further analysis would be required to draw any conclusions from this, as many of these eating events may have a low energy content.

In summary, the results of inference at the individual level has shown that in general there is a good match between the food groups consumed in an eating event and the subjective label for the most likely eating event type. This provides further confirmation that topic models can successfully be used to detect combinations of food groups reflective of different types of eating events. Furthermore, it has been demonstrated that the discovered eating event types provide a basis for further analysis at both the individual and population level. In particular, they can help to identify and investigate eating behaviour patterns for an individual. This could be extended to long term food diary data to identify stronger trends and links to health outcomes. Additionally, specific hypotheses at the population level, such as “skipping breakfast is linked to a lower daily energy intake” can be investigated using the discovered eating event types.

7.1.2 Investigating differences between weekdays and weekends

In addition to performing inference at the individual level the dataset can be explored by performing inference for different groups based on a variable of interest. The original method of document creation, using all items consumed at time t for any person on any day, is applied to each group. Visualising the results from inference on the groups can highlight differences between them. For example, the dataset can be grouped based on the day of the week variable, with all of the weekdays from Monday through Friday in one group and the weekend days, Saturday and Sunday in a second group.

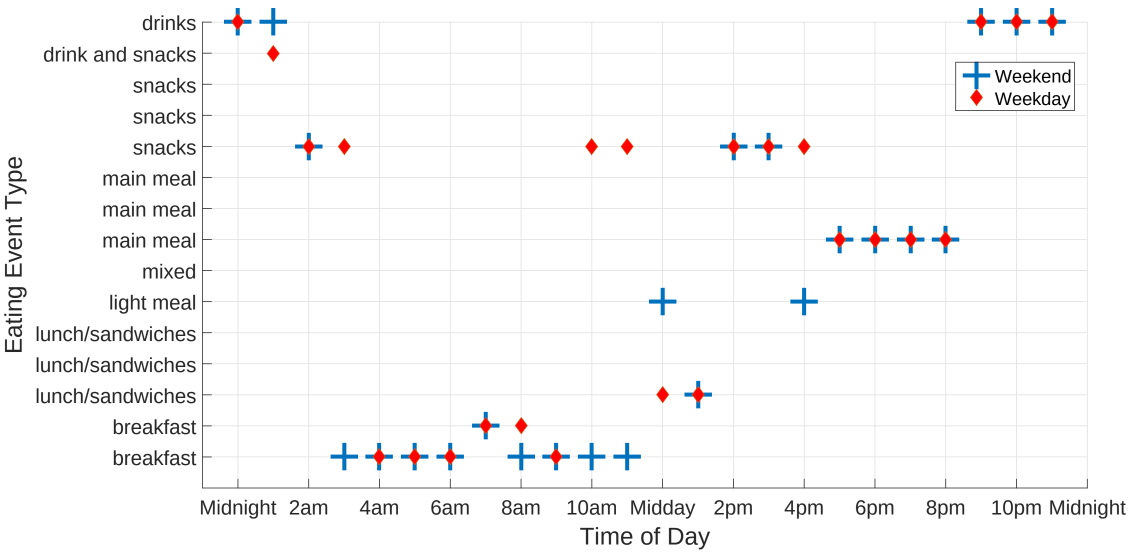


Figure 7.6: Most likely eating event types for each hour when inference is performed on groups of weekday and weekend days.

Figure 7.6 shows the most likely eating event type for all of the eating events in each hour, for weekdays and weekend days. This highlights that there are many similarities in the types of eating events that occur on both weekdays and weekend days. However, there are some differences which may be of interest to nutrition researchers. In particular, it can be seen that on weekend days the period of time for which breakfast is the most likely eating event type is longer than on weekdays. This could suggest that people have less of a routine at the weekend and this may in turn

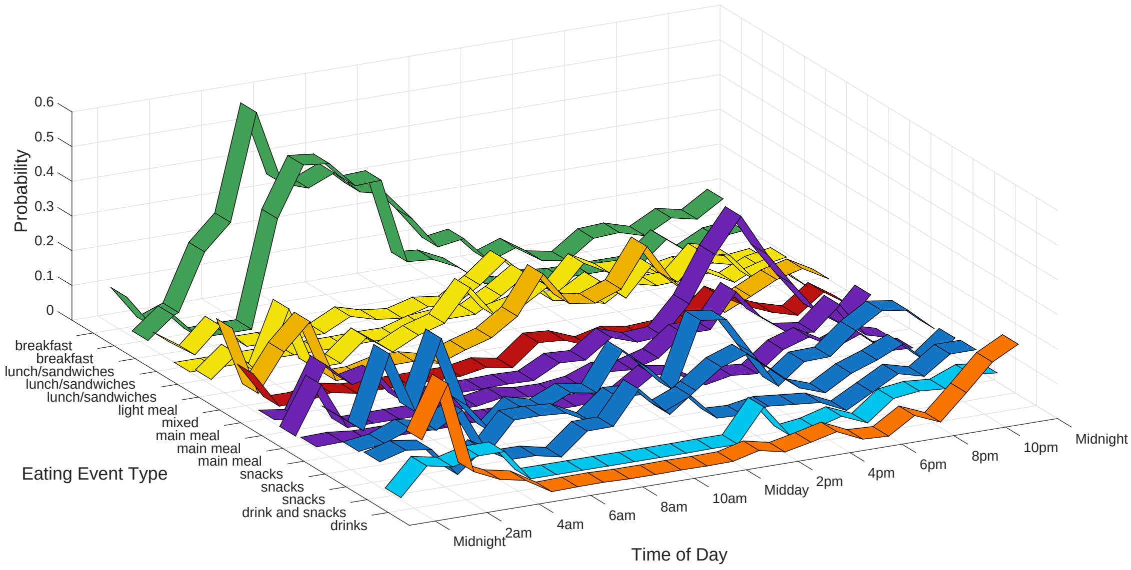


Figure 7.7: Probability of 15 eating event types at different times of day when inference is performed on weekend days.

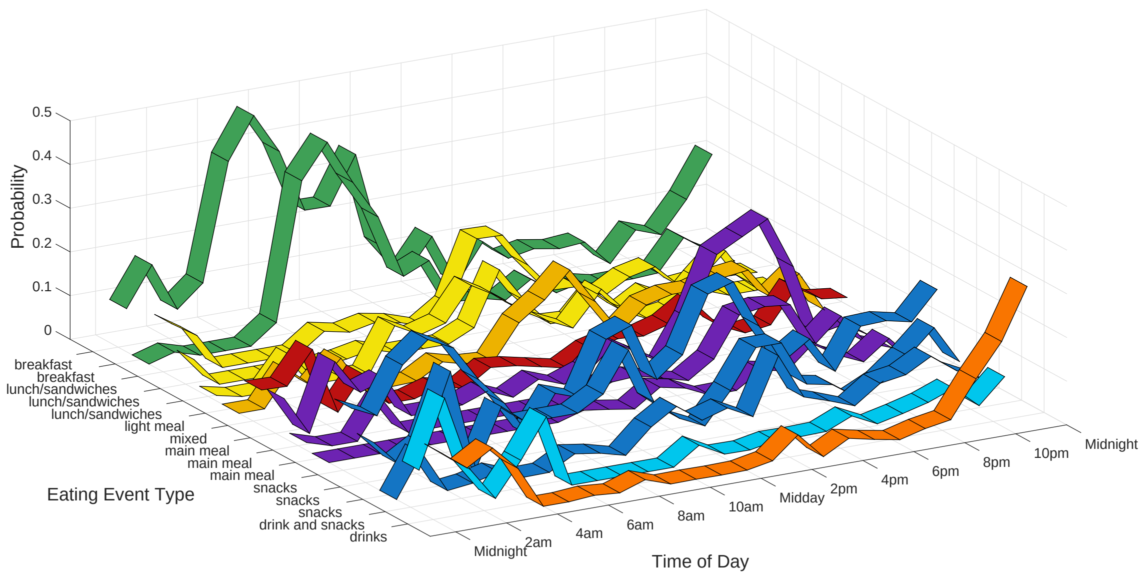


Figure 7.8: Probability of 15 eating event types at different times of day when inference is performed on weekday days.

impact what is consumed for the remainder of the day. Nutritionists could utilise these eating event types to investigate this at a higher level than has previously been possible when using the raw data directly [231].

One of the benefits of LDA is its probabilistic nature, that allows the uncertainty in the data to be modelled. For example, the model can accommodate the fact that not everyone eats the same food at the same time for breakfast. The probabilities of the eating event types found from performing inference on weekend days and weekdays can be visualised at different times of day, as shown in figures 7.7 and 7.8 respectively.

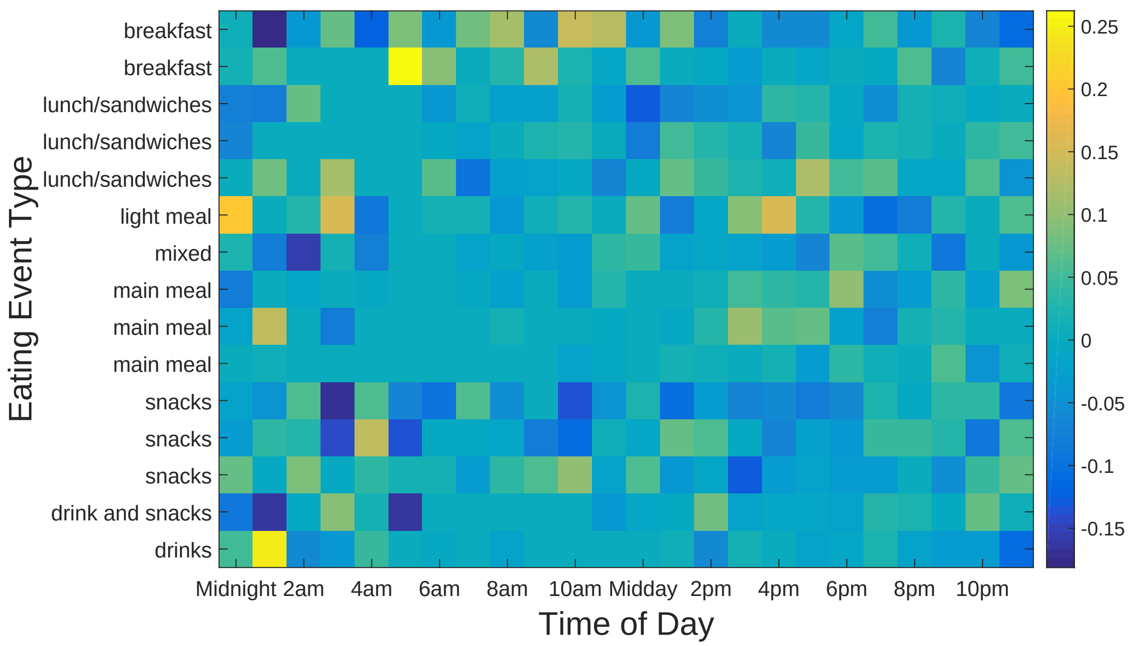


Figure 7.9: Differences between probabilities of 15 eating event types at different times of day. Positive values (yellow and green) indicate higher probability for weekend and negative values (blue) for weekdays.

Comparing these graphs provides a more detailed insight into the differences that were identified in figure 7.6. For example, on weekdays there are two strong peaks of the snack event type either side of peaks for the lunch/sandwiches and light meal event types. Whereas, at the weekend there is more uncertainty between all of the event types around midday. Additionally, more detail between the probability of the two different breakfast eating event types can be seen on these graphs, than is summarised by visualising only the most likely eating event types.

In order to quantify the observed variation in probabilities between weekdays and weekends the differences between the two groups are calculated. These are visualised as a matrix of eating event types at different times, as shown in figure 7.9. A positive value, green and yellow squares, indicates that this eating event type is more likely at this time of day on a weekend day. Vice versa, a negative value, blue squares, indicates that the eating event type is more likely at this time of day on a weekday. For example, at 5am one of the eating event types, labelled as breakfast, is a lot more probable on a weekend than on a weekday. Conversely, at the same time the eating event type drinks and snacks is more likely to occur on a weekday. The matrix visualisation shows that overall there is not much difference between the probabilities for each type of day. Moreover, the larger differences in probability tend to occur between midnight and 6am, during which period less eating events tend to occur, meaning there are fewer samples and hence each event will have a larger impact on the probability.

In summary, this section has demonstrated that performing inference on different groups can reveal patterns that are useful for further analysis. In particular, it has been shown that there are differences between the probabilities of eating event types that occur on weekday and weekend days. The eating event types discovered by a topic model can be used by nutritionists as a basis to investigate when and where these food group combinations are consumed, their average energy and other variables of interest. Understanding the patterns of consumption for the UK population on average can help to influence policy on nutrition and dietary advice.

7.2 Validating application of topic models to nutrition data

It has been shown that topic models can be applied to food diary data from the NDNS RP dataset to identify combinations of food groups that are representative of different types of eating events. The results have been qualitatively and quantitatively evaluated and demonstrate that this method is successful. However, it is necessary to both confirm that this method extends to other datasets and validate the results against other approaches. There is no gold standard within the nutrition research community for identifying eating event types, as discussed in section 3.3.3. The discovered eating events are validated against the results of other techniques [202] and subjective labels given by participants [183].

Several datasets were considered for the purposes of validation. The United States National Health and Nutrition Examination Survey [232] has datasets available for every year from 1999 to present. However, the details of the survey change each year, only 2 days of food diary data are collected for each participant using the 24 hour dietary recall method and the food groups are not given explicitly in the dataset. The European Prospective Investigation into Cancer and Nutrition study [180] has datasets from 10 different countries, however there is only one 24 hour dietary recall for each participant. Using a different method of data collection, dietary recall instead of food diaries, means that the comparison between the results may be less valid and hence these datasets were not selected for the validation.

The UK National Diet and Nutrition Survey 2000 (NDNS 2000) [233] and the Irish National Adult Nutrition Survey (NANS) [183] were chosen for use as validation sets. Details of these datasets are provided in sections 7.2.1 and 7.2.2 respectively. Both of these datasets included four or seven day food diary data, making comparisons across the datasets easier. Moreover, these datasets have previously had alternative manual approaches to identifying eating event types applied to them, hence enabling direct comparisons between the results of the different methods. Whilst there are similarities between the datasets, there are also differences in the food lists and coding of food groups, particularly for the NANS dataset. Therefore, they can both be used for validation but direct comparisons between them are more difficult.

7.2.1 Validation using UK National Diet and Nutrition Survey 2000 dataset

The UK National Diet and Nutrition Survey 2000 is the predecessor to the NDNS RP dataset used for the initial work. It was a cross sectional survey of a representative sample of the UK adult (aged 19 - 64 years) population living in private households. The survey investigated the food consumption, nutrient intake and nutritional status of participants over a 12-month period between 2000 and 2001. This included the completion of 7-day weighed dietary records of food intake. Participants were provided with accurately calibrated weighing scales and instructions on how to weigh food, drinks and leftovers. Full details of how the survey was conducted and all of the data collected is available from the UK dataservice and was accessed under usage number 87742 [233]. From the full eligible sample in the survey, 1724 participants completed the full seven day weighed food diary. The data from the diaries were coded by trained individuals using a list of food codes for 6000 items, provided by the survey organisers. Composite items were split into individual components, for example a cup of tea may be recorded as tea infusion, milk and sugar [234].

7.2.1.1 Methods for comparing manual rule based and topic model labels

A manual rule based classification method for identifying eating events as meals, snacks or drinks has recently been applied to this dataset [202]. This method involved allocating all of the food groups in the NDNS 2000 dataset to the most relevant list selected from meal, snack and drink, based on expert knowledge. The lists are given in appendix B.3 for reference. Dietary supplements were not included in this analysis. In the published work eating events were considered to be all items consumed within every 60 minute period, but this method was also applied for eating events defined as all items consumed at a unique time. The results based on eating events at unique times were used to facilitate comparison with the work described in this thesis. The following rules were then applied to label an eating event as a meal, snack or drink:

Meal All items from meal list OR
More than one item and at least one item from meal list EXCEPT

Two items only, one each from meal and snack lists (e.g. bread and butter)

Snack All items from snack list OR
Two items only, one each from meal and snack lists

Drink All items from drink list OR
Two items only, one each from drink and snack lists (e.g. sugar in coffee)

Although the analysis presented in [202] excludes several of the participants due to missing non-dietary data, the full sample was used in this work as all dietary data was complete. The manual rule based labels, assigned using the above method, for all participants was provided by Dr Laura Johnson, University of Bristol, a collaborator for this PhD thesis. In order to be able to draw comparisons and generalise the results across the datasets investigated, the same approach was taken to applying topic models to this dataset as used for the NDNS RP dataset. Documents were constructed from the items consumed in all of the eating events that occurred at every unique time t, as recorded to the nearest minute, for any participant on any day, giving a total of 1148 documents. The vocabulary consists of the 57 food groups found in the NDNS 2000 dataset, which are broadly the same as those in the NDNS RP dataset and are listed in appendix B.2 for reference.

The C implementation of variational inference for LDA from David Blei [51] was also used for this dataset for consistency. A ten-fold cross validation was conducted using the method described in 6.2.1, with 114 documents in each fold. The average perplexity was calculated for a range of eating event types from 5 to 50 and is shown by blue crosses in figure 7.10. It can be seen that the perplexity is lowest for between

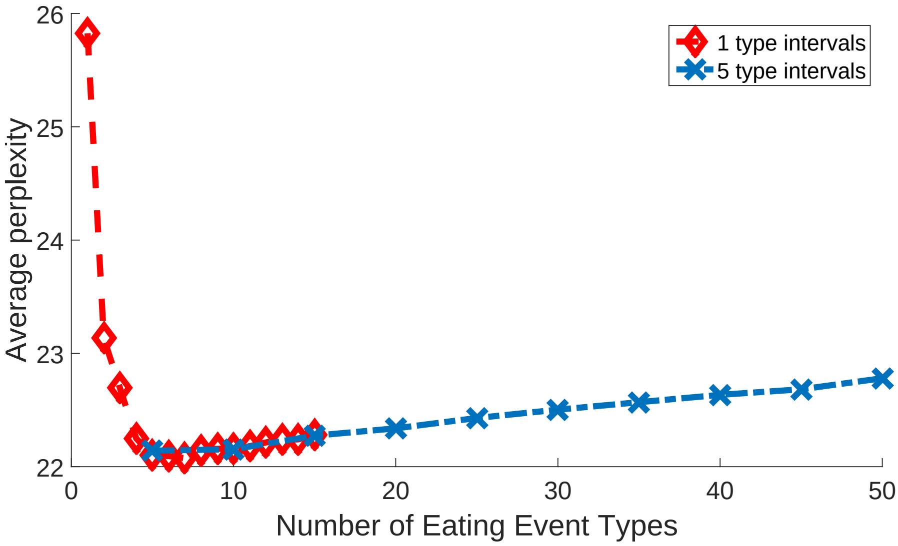


Figure 7.10: Graph showing how the average perplexity varies with the number of eating event types for the NDNS 2000 dataset.

5 and 10 eating event types, but that it does not increase much for the full range of numbers. This suggests that the model generalises to previously unseen data best when using 5 or 10 eating event types, however the number of types does not have a strong effect, as for the previous dataset. A more detailed cross validation was also conducted for between 1 and 15 eating event types at 1 type intervals, shown by red diamonds in figure 7.10. This confirms that between 5 and 10 types gives the lowest perplexity, with the minimum at 7 types.

In addition a qualitative analysis was also conducted for the dataset. The results of estimating models using 5, 10 and 15 eating event types were visualised as lists of the top food groups with font size proportional to the probability of the food group, given in appendices C.3 - C.5. These were evaluated for their semantic coherence and given subjective labels using the same types as for the previous dataset, as summarised in table 7.1. All of the eating event types found were considered to be semantically coherent, however the model with 5 eating event types is a limited representation as there are no event types associated with lighter meals.

For comparison with the manual rule-based classification method it is necessary to map the eating event types given in table 7.1 to the three simple types used for the rules. Therefore, breakfast, lunch/sandwiches, light meal and main meal were all mapped to the label meal; whilst drinks and snacks remained with the same labels. The NDNS 2000 dataset was restructured into documents that were created based on single eating events, in the same way as described in section 7.1.1. These documents

Table 7.1: Mappings of eating event type numbers to subjective labels for models estimated using the NDNS 2000 dataset for 5, 10 and 15 types of eating events.

Type #	Eating Event Type Label		
1	Snacks	Main meal	Snacks
2	Drinks	Breakfast	Breakfast
3	Main meal	Main meal	Snacks
4	Breakfast	Snacks	Light meal
5	Drinks	Lunch/sandwiches	Light meal
6		Light meal	Snacks
7		Lunch/sandwiches	Main meal
8		Drinks	Lunch/sandwiches
9		Drinks	Lunch/sandwiches
10		Snacks	Drinks
11			Snacks
12			Breakfast
13			Breakfast
14			Drinks
15			Lunch/sandwiches

were used to perform inference and determine the most likely eating event type for every eating event. This was completed for the models estimated using 5, 10 and 15 eating event types. The simple eating event type labels (meal, snack, drink) for each eating event could then be directly compared with the label given to the same eating event using the manual rule-based method.

7.2.1.2 Results and discussion of comparison between manual rule based and topic model labels

Table 7.2 shows the percentage of eating events that were given the same label by both the manual method and the topic model approach, using the mapped simple labels. The model with 15 original eating event types has the highest similarity to the manual rule-based method. This suggests that although the eating event types are mapped onto only three labels, starting with a large number of types when estimating the model can still provide benefit.

Table 7.2: Percentage of matched labels between rule based method and labels assigned by models for 5, 10 and 15 eating event types for the NDNS 2000 dataset.

Percentage of Matched Eating Event Types		
5 Type Model	10 Type Model	15 Type Model
67.7%	66.3%	70.6%

Table 7.3: Detailed break down of fraction of eating events for each manual rule based label that matched each mapped simple label for the most likely eating event type assigned by the model, for models estimated with 5, 10 and 15 eating event types using the NDNS 2000 dataset.

		5 Type Model			**10 Type Model**			**15 Type Model**		
		Meal	Drink	Snack	Meal	Drink	Snack	Meal	Drink	Snack
Manual Label	Meal	0.84	0.01	0.16	0.92	0.01	0.07	0.91	0.08	0.01
	Drink	0.23	0.32	0.45	0.34	0.44	0.22	0.25	0.58	0.17
	Snack	0.12	0.10	0.78	0.19	0.18	0.63	0.25	0.10	0.66

For each of the models, table 7.3 shows a matrix with the fraction of eating events for each manual label that matched each of the mapped simple labels for the most likely eating event type assigned by the model. For example, for all of the eating events with the manual rule based label of 'Meal', the fraction of labels assigned using the topic model with 5 different eating event types are 0.84 for 'Meal', 0.01

for ‘Drink’ and 0.16 for ‘Snack’, as shown by the first row of the first matrix. These results highlight where the differences occurred between the two labelling methods. For all of the models it can be seen that meals are the most likely to be given the same label by both methods. There is a much larger disagreement for the drink label, but the agreement for this improves as the number of eating event types in the model increases. This is likely to be because the eating event types discovered become more specific when a larger number are found. The agreement for the snack label seems to vary and is highest for the model with 5 eating event types.

However, the manual rule-based method would not be considered a gold standard. When investigating the details of the eating events that are given different labels it can be seen that identifying which method is correct is very subjective and not always possible to determine. Table 7.4 summarises some examples of eating events that are given different labels by each method. These are discussed in further detail here.

In general, it was observed that eating events with only one item often cause

Table 7.4: Examples of items in different eating events to highlight the difference in manual and model label assignments.

#	Food Items	Manual Label	Model Label
1	Chicken Arrabbiata ready meal	Snack	Meal
2	Semi-skimmed milk	Drink	Meal
3	Coca cola Chocolate bar	Drink	Snack
4	Instant coffee powder Water as diluent Milk Sugar	Snack	Drink or Meal
5	Medicine Water	Drink	Snack
6	Cream crackers Brie Cream cheese	Snack	Meal
7	Bread (toasted) Marmite Instant coffee powder Water as diluent Milk Sugar	Meal	Snack or Drink
8	Mineral water Canned vegetable soup	Meal	Snack or Drink

confusion for both methods. In particular, the manual rule-based method assumes that a meal will have at least two items and therefore incorrectly labels some eating events which would subjectively be considered a meal, such as ready meals, pizza or a 'Big Mac'. Example 1 in table 7.4 shows a ready meal, which is labelled as a snack by the manual method. In these cases, the topic model approach performs better as it labels the eating event as a meal. However, in other instances of one item eating events, notably drinks, the topic model will give a label of meal, although drink is more appropriate, as demonstrated by example 2.

The manual rule-based method will label an eating event with two items, where one is a drink item and the other a snack item, as a drink, in order to account for people who have sugar in their hot drink. However, this does not cover cases where a hot drink is listed as several separate components, such as in example 4, where the manual label is incorrectly given as snack. For this example, the topic model approach correctly labels this eating event as a drink when using the 5 eating event type model but incorrectly gives the label meal for the models with 10 or 15 eating event types. Moreover, this rule also misses cases where a drink and snack are consumed, such as in example 3. This is correctly labelled by the topic model method.

Other eating events are less obviously of a particular type and therefore it is not possible to say which label is correct. Examples 5, 6 and 7 in table 7.4 highlight some common types of eating events that have this problem. The addition of medicine to the eating event in example 5 makes it ambiguous as to whether it is still just a drink or should be considered a snack. It is possible that removing all medicine items from the dataset would be appropriate for this analysis. However a decision would then need to be made as to whether the associated water was relevant.

Example 6 is also ambiguous as some people would eat this combination of foods as a snack, whilst others would consider it to be a lunchtime meal. Therefore, giving an eating event like this a label is difficult and not necessarily that meaningful. Similarly for example 7, some people might have marmite on toast and coffee for breakfast, whereas another person might have this as an evening snack. However, for this eating event the models with 5 and 10 eating event types label this as a drink, which is incorrect. In general, the models with less types appear to be more likely to label eating events that are meals or snacks as drinks.

Finally, example 8 demonstrates a problem with eating events where soup is consumed and there are not many items in total. These are often labelled incorrectly by the topic model approach. This may be because the food group that soup is part of is 'miscellaneous' and this food group also contains savoury sauces, condiments and dry beverages, such as hot chocolate powder. Therefore, it is not discriminatory enough and causes confusion in the model. Whereas the rule based approach utilises the 115 sub food groups and hence soups are explicitly assigned to the meal list. Redefining the food groups used specifically for applying topic models to find eating event types could improve the results by avoiding such ambiguities but still maintaining a small vocabulary.

In summary, this section has shown that the results found in chapter 6 extend to a different dataset. Topic models have successfully been applied to the NDNS 2000 dataset to find combinations of food groups representative of different types of eating events. The method was easily applied to this new dataset, even though the vocabulary was different, the food diary records were for seven days rather than four and the survey only included adults aged 19 - 64 years. Furthermore, it has been demonstrated that the eating event types found by the topic model are comparable to those found using a manual rule-based approach in up to 71% of cases. Using a model with a larger number of topics gives a higher agreement between eating events labelled as drink because the discovered eating event types are more specific. For some eating events the topic model results are correct when the rule-based method fails.

Finally, the subjective nature of labelling eating event types is shown to not be applicable for all situations as often the same items will be consumed but considered different types of events. This suggests that using the eating event types found directly, rather than applying labels to them could be more insightful when using these results for further analysis. However, this can cause problems for nutritionists when trying to communicate guidelines for policies to a lay audience as it is normal to use these labels for communicating such ideas.

7.2.2 Validation using Irish National Adult Nutrition Survey dataset

The Irish National Adult Nutrition Survey (NANS) was conducted by the Irish Universities Nutrition Alliance (IUNA) to investigate food consumption, lifestyle and health of a representative sample of the population of the Republic of Ireland. The survey ran from 2008 - 2010 and recruited 1500 free-living adults aged 18 - 90 years, excluding women who were pregnant or breastfeeding. Participants completed a four-day semi-weighed food diary, weighing food items when possible, using manufacturers information or a food atlas to estimate weights otherwise. Participants were asked to define the type of each eating event and this was coded at the time of input to the database. Full details of the survey are available in the summary report [183]. Data was made available through collaboration with Dr Eileen Gibney, University College Dublin, a member of IUNA.

7.2.2.1 Methods for comparing participant defined and topic model labels

The data from the food diaries for all 1500 participants were coded using the WISP (Tinuviel Software), using a total of 2552 items. The data was recorded at the item level in an SPSS Statistics (IBM) database and the relevant variables were extracted for processing in Matlab (v2014b, The Mathworks Inc.). The same method of applying topic models as used for both NDNS datasets was utilised for this dataset for consistency. Documents were constructed by identifying every unique time, to the nearest minute as recorded, in the dataset and selecting all of the eating events that occurred at this time. The food groups for every item consumed during each of the eating events were listed to create the document. The dataset contains 68 food groups and a reduced set of 19 food groups. The full set was used for creating the documents (listed in appendix B.2 for reference) as this was most similar to the NDNS food groups, however there are differences in the food groupings, which are highlighted in section 7.2.3. Overall, there were 539 documents for the full dataset.

The same implementation of variational inference for LDA was used as for the previous datasets for consistency. A ten-fold cross validation was conducted using the method described in 6.2.1, with 53 documents in each fold. The average perplexity for 5 to 50 eating event types, at intervals of 5 was calculated across all folds, as shown by blue crosses in figure 7.11. It can be seen that the perplexity is lowest for 5 eating event types but that it does not increase much across the full range. This is similar to the previous datasets and suggests that the model generalises to previously unseen

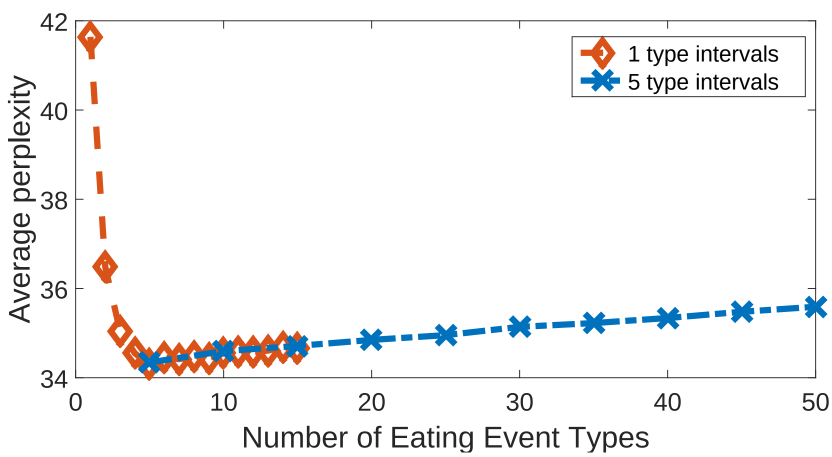


Figure 7.11: Graph showing how the average perplexity varies with the number of eating event types for the NANS dataset.

data best when using 5 eating event types, however it is not strongly affected by the number of eating event types. A more detailed cross validation was conducted for 1 to 15 eating event types and is shown by red diamonds in figure 7.11. This confirms that the lowest average perplexity value is at 5 types.

In addition to quantitative evaluation, it is also important to qualitatively analyse the results. Therefore, models were estimated using 5, 10 and 15 eating event types and the results were visualised as lists of the top food groups with font size proportional to the probability of the food group, given in appendices C.6 - C.8. These were evaluated for their semantic coherence and given subjective labels using the same types as for the previous datasets, as summarised in table 7.5. All of the eating event types found were considered to be semantically coherent. However the model with 5 eating event types has less variety, in particular meat based food groups only appear once in the top food groups of any eating event type. The results of the model with 15 types was judged by nutrition experts to be the most useful, particularly because it contained an eating event type for drink only events.

The label for the type of each eating event as defined by the participants was coded to be one of eleven options (the meal type labels listed in column two of table 7.6). These are different in nature to the labels used to subjectively name the eating event types found by a topic model because they focus more on the time of day. For example, there are four types of snacks that only differ by the time of day they are

Table 7.5: Mappings of eating event type numbers to subjective labels for models estimated using the NANS dataset for 5, 10 and 15 types of eating events.

Type #	**Eating Event Type Label**		
1	Breakfast	Drinks and snacks	Lunch/sandwiches
2	Light meal	Light meal	Lunch/sandwiches
3	Drinks and snacks	Main meal	Lunch/sandwiches
4	Main meal	Breakfast	Drinks and snacks
5	Drinks and snacks	Main meal	Main meal
6		Drinks and snacks	Breakfast
7		Lunch/sandwiches	Main meal
8		Lunch/sandwiches	Lunch/sandwiches
9		Breakfast	Lunch/sandwiches
10		Main meal	Drinks (alcoholic)
11			Main meal
12			Breakfast
13			Light meal
14			Main meal
15			Light meal

Table 7.6: Mapping from labels defined by participants (meal type labels) and subjective labels assigned to discovered eating event types (eating event type labels) to a common set of simple labels to enable comparison

Simple Labels	Meal Type Labels	Eating Event Type Labels
Breakfast	Breakfast	Breakfast
Light meal	Lunch light meal Evening light meal	Light meal Lunch/sandwiches
Main meal	Lunch main meal Evening main meal	Main meal
Drink / Snack	Morning snack Afternoon snack Evening snack Night snack Non-alcoholic beverage Alcoholic beverage	Snacks Drinks and snacks Drinks (alcoholic)

associated with. In order to make comparisons between the two approaches, a reduced set of simple labels was identified, to which the labels used in the two approaches could be mapped. The mappings between these different sets of labels are explicitly defined in table 7.6. Drinks and snacks are combined as only one discovered eating event type represented only alcoholic drinks.

In order to compare the labels given to each eating event, documents were created for each single eating event in the NANS dataset using the method described in section 7.1.1. Inference was performed on these documents using each of the three estimated models, with 5, 10 and 15 eating event types. The most likely type was determined for every eating event. The mappings for the simple labels, given in table 7.6, were then applied to the results from the models and the 'meal type' variable in the original dataset. This allowed the simple labels to be directly compared for each eating event to quantify the similarities between the model results and the participant defined labels. Table 7.7 shows the percentage of eating events that were mapped to the same simple label. The model with 5 eating event types has the highest similarity to the participant defined labels when mapped to the simple label set. The percentage of matches decreases with the increase of the number of eating event types.

For each of the models, table 7.8 shows a matching matrix detailing the fraction of eating events for each simple label (mapped from the participant defined labels) that matched the simple labels (mapped from the most likely eating event type assigned by the model). These results highlight that the majority of the decrease in overall percentage of matched types as the number of eating event types in the model

Table 7.7: Percentage of matched labels between participant defined labels and labels assigned by models for 5, 10 and 15 eating event types for the NANS dataset.

Percentage of Matched Eating Event Types		
5 Type Model	10 Type Model	15 Type Model
71.6%	60.9%	56.5%

Table 7.8: Detailed break down of fraction of eating events for each simple label (mapped from participant defined labels) that matched each simple label (mapped from the subjective eating event type labels) for the most likely eating event type assigned by the model, for models estimated with 5, 10 and 15 eating event types using the NANS dataset.

		5 Type Model			
		Breakfast	Light meal	Main meal	Drink/snack
Meal Type Label	Breakfast	0.78	0.07	0.00	0.15
	Light meal	0.10	0.61	0.07	0.21
	Main meal	0.02	0.16	0.72	0.10
	Drink/snack	0.18	0.07	0.03	0.72

		10 Type Model			
		Breakfast	Light meal	Main meal	Drink/snack
Meal Type Label	Breakfast	0.85	0.07	0.16	0.06
	Light meal	0.18	0.58	0.10	0.14
	Main meal	0.03	0.28	0.63	0.06
	Drink/snack	0.25	0.09	0.17	0.48

		15 Type Model			
		Breakfast	Light meal	Main meal	Drink/snack
Meal Type Label	Breakfast	0.73	0.24	0.02	0.02
	Light meal	0.07	0.78	0.10	0.04
	Main meal	0.01	0.35	0.60	0.04
	Drink/snack	0.22	0.22	0.18	0.38

increases is related to those eating events defined by participants as drinks or snacks. As the number of eating event types discovered increases they become more specific. For this dataset, the more specific eating event types found were more often subjectively labelled as a type of meal, especially lunch or light meal than as a snack or drink. These labels do not match as well with how participants define their eating events. In particular, an eating event defined as a snack by a participant can have similar food groups to part of a meal, especially eating events with few items.

Several of the more specific eating event types in the 15 type model that were labelled as lunch or light meal contain food groups related to bread, tea, coffee, milk, sugar and spreads within the top food groups of the eating event type. The subjective labels are assigned based on the distinguishing food groups, such as cheese, bacon & ham, soups and salad, which generally have a lower probability within the eating event type. The eating event type discovered labelled as 'snacks and drinks' also contains tea and milk within the top food groups but is distinguished by the chocolate confectionery, savoury snacks and cakes, pastries & buns food groups. This means that eating events that only have food groups such as bread, spread and tea have a higher chance of being assigned a most likely eating event type with a subjective label of light meal or lunch. This highlights the problem of subjective, cultural based labels as in reality people eat these same combinations of food groups at many different times but associate different labels with them. For example, if a person has toast and tea for breakfast or a night time snack, or a sandwich and tea for lunch, they are consuming very similar food groups but using different labels.

Due to the subjective nature of the labels given to eating events by participants and to the results from topic models by researchers, only matches and differences can be quantified, without stating that one is necessarily more correct than the other. Investigating the components of the eating events given different labels gives further insight into why these differences occur. Table 7.9 gives some examples of eating events which were given different labels. These are discussed in more detail below.

In the NANS dataset some items, such as milk and tea, are given average portion sizes. This means that if someone has a large mug of tea or two cups of tea in one eating event then this item will be listed more than once. This artificially inflates the representation of these items when creating documents at the eating event level, affecting inference of the most likely type of eating event. Example 1 in table 7.9 demonstrates this effect, as the distinguishing food group between a snack and a light meal, associated with the tuna food item, only accounts for 11% of the document created for this eating event. In addition, cream crackers are assigned to the biscuit food group and hence will be associated by the model with a snack eating event type rather than a light meal.

It was shown that, for the 10 and 15 type models, the majority of differences are for eating events with participant defined labels mapped to the drink/snack simple label. Eating events with only one or two items often cause differences in labels, especially for the 15 type model, as highlighted by examples 2 to 4 in table 7.9. Each of these events were labelled by participants as drink/snack but assigned to light meal, breakfast or main meal by the 15 type model. The food groups for the items in these eating events do not occur in the top food groups for the eating event types

Table 7.9: Examples of items in eating events with different labels assigned by participant and model approaches.

#	Food items	Meal Type Label	Model Label
1	Cream crackers Whole milk x 2 Tea, black infusion x 2 Water Tuna, canned in brine Crisps Mayonnaise	Light meal	Drink/snack
2	Apple	Drink/snack	Main meal
3	Coffee, infusion Whole milk	Drink/snack	Light meal
4	Cranberry juice	Drink/snack	Breakfast
5	Iced fairy cake White bread Whole milk Tea, black infusion Butter Chicken, roasted	Drink/snack	Light meal
6	Brown soda bread Marmalade Whole milk Tea, black infusion	Light meal	Breakfast
7	Brown bread Hash browns Whole milk Eggs, fried Coffee, instant Butter Bacon Pork sausages	Breakfast	Light meal
8	Crumpets Eggs, boiled Maple syrup Coke	Breakfast	Drink/snack
9	Beef stew Digestive biscuits Semi-skimmed milk Tea, black infusion	Main meal	Drink/snack

mapped to the simple drink/snack label for this model. Rather, they occur as part of other types of eating events, for example fruit can be eaten as part of a main meal. Therefore, because these eating events only have one or two items there is no data that the model associates with the drink/snack eating event type. Conversely, the eating event types discovered by the 5 type model are less specific, hence those types mapped to drink/snack include coffee and fruit. Therefore, the model can associate these eating events with the drink/snack types.

Example 5 in table 7.9 shows that for some eating events the label assigned by the model is arguably more realistic than the participant defined drink/snack label in terms of the content of the eating event. Another participant eating the same items could consider the eating event to be a light meal. This ambiguity due to the nature of the subjective labels also occurs for other eating event types, as shown in examples 6 and 7. Example 6 demonstrates that although the combination of tea, bread and marmalade is often consumed as breakfast this participant had the same items as a light meal. Similarly, in example 7 a traditional cooked breakfast is represented, however in terms of food groups it is clear that this is more like a light meal, as suggested by the model results.

Certain food groups that contain a wide range of different items can cause differences in labels, such as example 8. Coke and maple syrup belong to food groups more generally associated with snacks than breakfast, therefore the model has assigned the most likely type to be snack/drink for this eating event. Furthermore, the NANS dataset contains some items for which the nutrition values are known at a recipe level and hence are not split into their composite parts. These items are only listed as one food group in a document, rather than several which does not accurately represent the eating event. This is shown in example 8, where the beef stew is the main part of the meal but only has one item compared to the tea and biscuits, which require three items. Therefore the model bases the assignment to the most likely eating event type on the tea and biscuits rather than beef stew and considers it to be a snack instead of a meal.

All datasets contain some inconsistencies, some of the inconsistencies in the NANS dataset are clearly incorrect. For example, one participant who had porridge, bananas, pears, honey and coffee for breakfast every morning has this eating event labelled as a non-alcoholic beverage for day 3, which is clearly a mistake. Other inconsistencies cannot be considered definitively incorrect without being able to refer to the original food diary records. For example, one participant had porridge, pasta with tomato sauce and apples for their evening meal. This does not seem like a common combination, however it could be a valid entry.

In summary, this section has shown that topic models can be successfully applied to the NANS dataset to find combinations of food groups representative of different eating event types. The subjective labels given to eating event types found are comparable to the subjective labels assigned by participants in up to 72% of cases. In contrast, to the results for the NDNS 2000 dataset where as the number of eating event types discovered increases the agreement between labels increases, for the NANS dataset the agreement decreases. This decrease in agreement primarily occurs for drink/snack eating event types. A model with a larger number of eating event types discovers more specific combinations of food groups. For the NDNS 2000 dataset this meant that a more specific drink eating event type was discovered that increased the number of matches for the drink label. However, for the NANS dataset having less specific eating event types gives a better agreement with participant labels because it reflects the increased variation from using participant defined labels, rather than standardised rule based labels.

The coding strategies vary a lot more between NANS and NDNS 2000 than between the two NDNS datasets because they are run by completely separate groups. In particular, the nature of the food groups varies in terms of what level items are grouped at e.g. whether tea, coffee and water are all in one group or in three separate groups. The use of composite foods i.e. whether a sandwich is listed as one item or all of the component parts and average portion sizes are different. These underlying coding strategies affect the resulting eating event types discovered by a topic model because they are directly linked with the vocabulary and document structure. For example, if average portion sizes are used then the corresponding food groups will be repeated if a participant has a double portion. These differences have a more noticeable effect when performing inference on documents representing single eating events, compared to the documents used to estimate the model, because they are much smaller and hence a repetition of one item will be more significant in relative terms.

Finally, as for the NDNS 2000 dataset, the results highlighted that using subjective labels for eating events is not always applicable. The same combination of food groups may be consumed but given different labels based on cultural norms. This problem is compounded when both the participant based labels and labels assigned to the eating event types found by topic models are subjective. There is often no definitive answer as to which is the ‘correct’ or ‘better’ label.

7.2.3 Using NDNS food groups with the NANS dataset

The are several differences between the 68 food groups used in the NANS dataset and those in the NDNS datasets. For example, in the NDNS datasets tea, coffee and water are one food group whereas in the NANS dataset these are three separate food groups. The opposite is true for alcoholic beverages, which are separated into several different groups in NDNS but all grouped together in NANS. These different groupings affect the results of using topic models as the food groups are used as the vocabulary. For example, if a dataset contains 100 tea, 100 coffee and 100 water items then this would result in the NDNS 'tea, coffee and water' food group being listed 300 times in total, whereas each of the separate corresponding NANS food groups would be listed 100 times each in total. This change in the underlying structure of the data affects how the model discovers the hidden structure of eating event types.

To investigate this further, the food items in the NANS dataset were reassigned to the NDNS RP food groups by a nutritionist. There were a few unclear items that required clarification before being assigned to the most appropriate group, including splitting items into component parts where necessary. For example, the food item 'cocoa powder, made up with semi-skimmed milk' in the NANS dataset was coded to be NDNS RP food groups 'Beverages dry weight' and 'Semi-skimmed milk'. The assigned groups for these additional items were verified and agreed by a nutrition expert, familiar with the NDNS dataset.

Using the 60 NDNS RP food groups as the vocabulary, documents were created for the NANS dataset. Documents were constructed from the assigned food groups of the items in all eating events at each unique time recorded in the dataset, to the nearest minute. Variational inference for LDA was used to conduct a ten-fold cross validation using the method described in 6.2.1, with 53 documents in each fold. The average perplexity for 5 to 50 eating event types, at intervals of 5 was calculated across all folds as shown by blue crosses in figure 7.12. This shows the perplexity is lowest for 5 eating event types, but does not increase much across the full range.

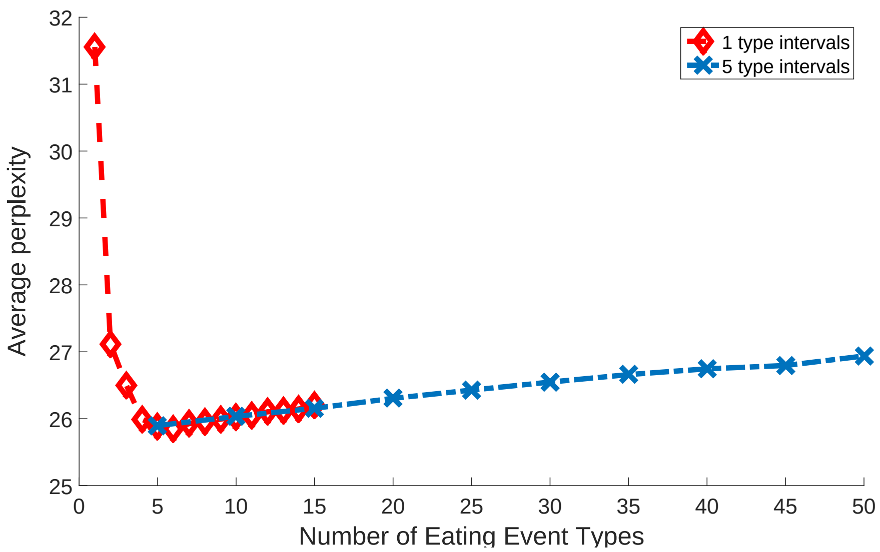


Figure 7.12: Graph showing how the average perplexity varies with the number of eating event types for the NANS dataset with the NDNS RP food groups.

A more detailed ten-fold cross validation was conducted and the average perplexity calculated for 1 to 15 eating event types, as shown by red diamonds in figure 7.12. This confirms that the lowest perplexity is for 5 or 6 eating event types. This is a similar trend to previous datasets and models, suggesting that the model generalises to previously unseen data best when using 5 eating event types, however the number of types does not have a large impact. Furthermore, comparing figure 7.12 to figure 7.11 highlights that the perplexity is lower across the whole range of number of eating event types when using the NDNS RP food group vocabulary as opposed to the original NANS food groups. This suggests that using the NDNS RP food groups creates models that are a better fit to the NANS dataset.

A qualitative analysis of the results was also carried out by visualising the top food groups for models estimated using 5,10 and 15 eating event types, given in appendices C.9 - C.11. All of the eating event types were considered to be semantically coherent. The models with larger numbers of types displayed more specificity in the types found in terms of food group combinations, as expected. All of the eating event types were given subjective labels using the same types as for previous datasets and these are detailed in table 7.10.

To further investigate the impact of using the NDNS RP food groups as the vocabulary for the NANS dataset, the same analysis of comparing labels as described in section 7.2.2.1 was conducted for the new models. Table 7.11 summarises the per-

Table 7.10: Mappings of eating event type numbers to subjective labels for models estimated using the NANS dataset with the NDNS RP food groups for 5, 10 and 15 types of eating events.

Type #	**Eating Event Type Label**		
1	Breakfast	Breakfast	Drinks and snacks
2	Main meal	Lunch/sandwiches	Main meal
3	Drinks and snacks	Drinks and snacks	Drinks and snacks
4	Drinks and snacks	Main meal	Main meal
5	Lunch/sandwiches	Main meal	Breakfast
6		Breakfast	Drinks and snacks
7		Lunch/sandwiches	Breakfast
8		Drinks and snacks	Light meal
9		Drinks and snacks	Drinks and snacks
10		Main meal	Drinks and snacks
11			Breakfast
12			Light meal
13			Main meal
14			Lunch/sandwiches
15			Lunch/sandwiches

Table 7.11: Percentage of matched labels between participant defined labels and labels assigned by models for 5, 10 and 15 eating event types for the NANS dataset with the NDNS RP food groups.

Percentage of Matched Eating Event Types		
5 Type Model	10 Type Model	15 Type Model
73.3%	69.5%	67.5%

Table 7.12: Detailed break down of fraction of eating events for each simple label (mapped from participant defined labels) that matched each simple label (mapped from the subjective eating event type labels) for the most likely eating event type assigned by the model, for models estimated with 5, 10 and 15 eating event types using the NANS dataset with the NDNS food groups.

		5 Type Model			
		Breakfast	Light meal	Main meal	Drink/snack
Meal Type Label	Breakfast	0.61	0.09	0.01	0.29
	Light meal	0.06	0.61	0.14	0.20
	Main meal	0.01	0.10	0.80	0.10
	Drink/snack	0.07	0.08	0.03	0.81

		10 Type Model			
		Breakfast	Light meal	Main meal	Drink/snack
Meal Type Label	Breakfast	0.85	0.08	0.01	0.07
	Light meal	0.16	0.62	0.10	0.12
	Main meal	0.03	0.18	0.71	0.08
	Drink/snack	0.22	0.08	0.06	0.64

		15 Type Model			
		Breakfast	Light meal	Main meal	Drink/snack
Meal Type Label	Breakfast	0.77	0.12	0.02	0.09
	Light meal	0.12	0.61	0.16	0.11
	Main meal	0.02	0.21	0.70	0.07
	Drink/snack	0.23	0.10	0.03	0.64

centage of eating events for which the final mapped label was the same for models with 5, 10 and 15 event types. It can be seen that as for the original NANS dataset and vocabulary, the model with 5 eating event types has the highest similarity to the participant defined labels when mapped to the simple label set. The percentage of matches also decreases with the increase of the number of eating events. Overall the percentage of matches is higher using the NDNS RP food groups, compared with the original NANS food groups and the rate of decrease in the number of matches is smaller. This suggests that the structure of the data using these food groups enables the topic models to discover hidden eating event types that are more aligned with participant based labels.

A detailed break down of the fraction of matches for each simple label are given for each model in table 7.12. These highlight that there is a similar decrease in the fraction of matching drink/snack labels between the 5 and 10 type models as there was for the original NANS food groups but that there is no further decrease for the 15 type model. This suggests that less specific eating event types still align better with participant defined labels but that using the structure of the NDNS RP food groups improves the agreement between the more specific eating events and the participant labels. One explanation for this is that by grouping 'tea, coffee and water' into one food group it is only one vocabulary item in the eating event types it appears in and hence there are a larger number of top food groups that are more discriminatory. This is in contrast to the eating event types discovered when the original NANS food groups were used where there were a large number of eating event types dominated by different proportions of combinations of tea, milk, bread, spreads and sugar, which are not discriminatory.

7.3 Chapter summary

This chapter has shown that the resulting eating event types from applying topic models to nutrition data from the NDNS RP dataset can be utilised for further analysis. This can occur at the individual participant level, the population sample level or for specific groups of interest. In particular, it has been demonstrated that the content of eating events, in terms of the food groups consumed, and the subjective label for the assigned most likely eating event type are well matched in the majority of cases. Patterns in the eating event types for an individual can be found and these can be used in future work to investigate a person's eating habits. Moreover, it was shown that the eating event types found can be utilised to investigate specific hypotheses, using the example of how breakfast habits affect the eating events that occur for the rest of the day in the sample population. Inference can also be performed on different groups within the dataset, based on a variable of interest, to reveal patterns that can be used for further analysis. An example was shown that highlighted the differences between the probabilities of eating event types that occur on weekday and weekend days.

In addition to using the results from the NDNS RP dataset, it has been demonstrated that topic models can successfully be applied to other datasets containing food diary data. Food group combinations representative of eating event types were found for the NDNS 2000 dataset and the NANS dataset. These datasets both have alternative sources of labels for eating event types: a rule-based method for the NDNS 2000 dataset and participant defined labels in the NANS dataset. Analysis was carried out to investigate the number of matches between the different label sources. The results demonstrated that the subjective labels assigned to eating event types found by topic models matched the rule-based labels in up to 71% of cases and the participant defined labels for up to 72% of eating events.

Some of the incorrect labels given to eating events by the results of the topic model approach were identified to be linked to the choice of vocabulary. Some food groups are too ambiguous and contain a variety of food items that would be consumed during different types of eating events, for example the NANS food group 'biscuits including crackers'. Different coding strategies, including how composite items are listed, the level of food groups and the use of average portions affect the resulting eating event types discovered as they alter the underlying structure of the data. Determining a coding strategy for specific use with topic models for identifying eating event types could help to improve the results.

Eating events with a small number of items, such as a cup of tea or coffee (sometimes listed as up to four items if split into coffee, water, milk and sugar), a piece of fruit or a glass of juice are often assigned incorrect labels by both the topic model results and the manual rule based method. Only considering the food groups for these small events does not provide enough information to reliably determine the type of eating event. However, simply labelling all events with one or two items as drink/snack is not necessarily valid and still does not account for a cup of coffee with four components. One suggestion is to pre-process the data with rules based on more detailed information, such as total energy content of eating event and then estimate topic models using only data for eating events above a defined energy content.

Overall, the results have shown that using a model with a larger number of eating event types aligns better with rule-based labels. Conversely, models with fewer types give a better match to participant defined labels. This is because the larger the number of types, the more specific the combinations of food groups in each eating event type is. These results reflect the fact that subjective participant labels are more varied in nature in comparison with applying a consistent set of rules. The choice of number of eating events type to use should be based on the nutrition research question being addressed.

Exploring the eating events that were assigned different labels highlighted that although sometimes each method can clearly give an incorrect result, often it is not possible to decide which method is correct. Many eating events could be given a range of labels, which could all be considered correct but only depend on variables such as time of day, which do not directly change the nutrient content of the items eaten. For example, a cup of tea and a piece of toast could be eaten as breakfast, a light meal or a snack. In each instance, although the label is different, assuming the quantities are the same, the nutritional content of the eating event is the same.

The benefit of the eating event types found using the topic model approach is that they do not have to be labelled. They are found based only on what is recorded in the dataset and not biased by cultural norms. The resulting food group combinations could be used for analysis without ever applying subjective labels to them. For example, there is real potential for investigating routines and consistency in eating habits. A higher variability in eating event types may be indicative of chaotic eating and hence lead to overconsumption and obesity. Currently, nutrition researchers only consider the standard deviation of energy intake in different time slots over multiple days. Using the food group combinations found by topic models would enable this analysis to be extended to define similarity in both the energy intake and the food sources of the energy. The drawback of not labelling the eating event types is the challenge of how to disseminate the results found in a format that will be suitable for use in public policy creation and health initiatives.

In conclusion, for the purpose of investigating how people actually combine food groups together in eating events topic models are better than the rule-based approach. This is because topic models are a data driven approach and is therefore less constrained than the rule-based approach which cannot account for the full variability in the data. Moreover, topic models do not rely on the use of predefined labels, which can be biased by cultural norms, rather the focus is on the food groups involved. However, using topic models with only food groups as input has limitations. More detailed and useful information regarding the energy, weight and nutritional content of an eating event is lost. The approach is also affected by the choice of food groups used, which is not standardised within the nutrition research community.

Chapter 8

Conclusions and Outlook

This chapter highlights the conclusions of the research presented in this thesis and how they address the stated research questions. An outlook for how this research can be extended through future work is also given.

8.1 Conclusions

The aim of the research conducted for this thesis was to detect behaviour patterns in data collected by a residential healthcare monitoring system in order to obtain information of a greater quality than the raw sensor data. To address this aim three research questions were formulated that focus on specific challenges in this area. This section summarises the conclusions of the research related to each of these questions.

8.1.1 Detection of daily routines for healthcare monitoring

Can patterns in a person's daily activities that are representative of routines be detected for a novel, real world activity dataset?

Chapter 4 presented the development of a mobile phone logging application that was used to collect a real world dataset for two healthy volunteers with 16 days of activity data each. Latent Dirichlet Allocation (LDA) was applied to the daily activities in this novel dataset and the results demonstrated that the detected groups of activities are representative of routines.

An average accuracy of up to 80% was reported when comparing the discovered routines with the recorded ground truth labels for each participant. However for this dataset, a higher accuracy was associated with a higher perplexity (across all numbers of routines), which suggests that the model does not generalise as well to unseen data. This observation highlights the impact of the subjective nature of both ground truth labelling and the assignment of labels to detected routines. The semantic coherence of routines is as important, if not more, than how well they correspond to ground truth labels. In particular, a discovered group of activities may represent a valid routine for a user that was not considered when labelling the dataset.

Considering the semantic coherence of the discovered routines resulted in the selection of models with fewer routines than suggested by the quantitative measure of the perplexity. Using a larger number of routines gives a lower perplexity, meaning that the model can generalise to unseen data better, but this is because the discovered routines are strongly associated with single activities rather than groups that form a routine. The duration of discovered routines must be long enough to include a mixture of activities, in order for information of a greater quality than the input to be obtained.

In addition, it is recommended that the overall duration of the dataset used to estimate the model is of sufficient length to represent the general variability of a user's behaviour, in order that the model is less sensitive to rare or unseen activities. The absolute duration necessary will be dependent on the specific research question and use case. For example, an elderly person with a very stable routine will need a shorter dataset compared to a student with a chaotic routine. Furthermore, datasets should be collected over consecutive days to ensure there is no bias in the selection of recording days.

In conclusion, LDA can successfully be applied to a novel activity dataset to detect patterns representative of routines. The selection of the number of routines for a model can be guided by the measure of perplexity but the semantic coherence of the routines in the context of the application must not be lost. Validating detected routines against ground truth labels can give a good indication of performance. However, analysing the content of the routines directly can reveal more useful information beyond the restriction of subjective predefined labels.

8.1.2 Identification of changes in routines over time

Can changes in the structure of daily routines over time be detected?

Following on from the results of chapter 4, that showed LDA can detect routines in activity data, the application of dynamic topic models (DTM) to detect changes in the structure of daily routines over time was considered in chapter 5. The results demonstrated that changes over time in the probabilities of activities within discovered routines can successfully be identified for a simulated dataset.

The results revealed that the relationship between the duration of the activities used as input to the DTM and the number of routines in the model affect how easily a change can be identified. For the simulated dataset used in this research models with 5 routines and only using activities that occur during waking hours (i.e. no sleep activities of a duration greater than an hour) performed the best. A novel visualisation technique was developed that highlights activities that vary by at least 25% of the minimum probability of the activity occurring in the routine, assuming this is over 0.001. These criteria were shown to be successful at identifying relevant changes for the simulated dataset tested.

It was observed that not all of the changes in the simulated dataset were identified at the activity level. In other words, some changes were identified as a variation in the frequency of a routine, rather than the probability of an activity within routine(s). This is due to the variation in duration of activities that are modelled in the simulated dataset. A limitation of using DTMs is that the model and dataset parameters need to be tuned using expert knowledge and contextual information for different scenarios.

In conclusion, changes in the structure of daily routines over time can successfully be detected using dynamic topic models for simulated datasets. This preliminary result suggests there is potential for this method to be successful when applied to real activity data. However, this approach is affected by the model and dataset parameters and how they relate to each other and therefore these must be carefully selected for each application. Furthermore, the criteria used to highlight relevant changes will need to be adjusted to reflect significant real world healthcare criteria. For example, using a threshold of 0.001 for the minimum probability of activities considered to be changing may not reflect any useful diagnostic information for clinicians.

8.1.3 Understanding how foods are combined in eating events

Can the combinations of foods consumed together in different types of eating events be automatically detected?

Chapter 6 presented a novel method for applying LDA to nutrition data recorded in the form of food diaries. The results demonstrated that combinations of food groups that represent different types of eating events can automatically be detected using LDA. This method was shown to successfully detect eating event types for three different datasets in chapters 6 and 7.

When visualising the results of the eating event types found for the NDNS RP dataset it was noted by a nutrition expert that some food groups often appeared with a relatively large probability compared to others. It was hypothesised that this was due to the structure of the original dataset. Chapter 6 detailed a sensitivity analysis, revealing that altering the structure of the data to eliminate this artefact did not improve the performance of the topic models and hence the original structure should be kept.

Chapter 7 showed that the eating event types detected for the NDNS RP dataset can be utilised for further analysis. Performing inference for single eating events enables patterns in eating behaviours to be investigated at both the individual participant and population sample levels. Inference can also be performed on a group of data related to a variable of interest to investigate similarities and differences between the groups, such as weekdays and weekend days.

The subjective labels assigned to the eating event types discovered by LDA in the NDNS 2000 and NANS datasets were validated using alternative sources of labels. The results given in chapter 7 revealed that the subjective labels matched the rule-based labels in NDNS 2000 and participant defined labels in NANS for up to 71% and 72% of the eating events respectively. Overall, it was observed that a model with a larger number of eating event types aligns better with rule-based labels and conversely for participant defined labels. This reflects the fact that there is more variety in how participants assign labels to eating events.

Eating events with only one or two items are often assigned incorrect labels by both the LDA and rule based approaches. This is because only considering the food groups for these small events does not provide enough information to reliably determine the type of eating event. Only using food groups as input data for topic models is a limitation of this method. Data regarding the energy, weight and nutritional content of an eating event, that can reveal important information, is lost. The approach is also affected by the choice of food groups used, which is not standardised within the nutrition research community.

In conclusion, combinations of foods consumed together in different types of eating events can be automatically detected using LDA. The labels assigned to the discovered combinations have a good agreement with other methods for labelling eating event types. However, as all types of labelling are subjective in nature and biased by cultural norms it is proposed that the combinations of food groups could be used for further analysis without applying labels. The drawback to this suggestion is the challenge in disseminating the results to lay audiences in a useful manner.

8.2 Outlook

During the course of this research several avenues for future work were identified. These are highlighted in this section with suggestions for extending the work presented in this thesis.

8.2.1 Future work for detecting routines and changes over time

Analysing the results of applying LDA to the novel activity dataset collected highlighted limitations of how the model deals with previously unseen data. The recommended guideline previously published by Seiter et al. [124] to use at least 14 days of data for good performance was followed. However, it was observed that over the 16 days of data for each participant in the novel dataset some activities only occurred on one or two days. The performance of the model decreased significantly for days with these activities when a leave-one-day-out cross validation was performed. It was observed that the days with lower accuracies for participant two were weekend days, whereas all the other days in the dataset were weekdays.

One improvement would be to collect data on consecutive days for a longer period of time and ensure that both weekend and weekdays are well represented. Future work, in collaboration with clinical experts, should address the question of how much data should be used to estimate a model that represents a baseline for establishing a person's 'normal' routine. The amount of data required may vary depending on the healthcare context and the nature of an individual's lifestyle. For example, a student is likely to have a more chaotic routine than a parent with a small child or an older person with a chronic disease.

Another possible solution for previously unseen data is to use online LDA with an infinite vocabulary proposed by Zhai and Boyd-Graber [211]. This method enables the model to expand topics to include additional words by drawing topics from a Dirichlet process (DP) with a base distribution over an infinite vocabulary, instead of a finite Dirichlet distribution. The infinite vocabulary is implemented using a truncated ordered set that is updated to allow the vocabulary to dynamically expand and contract as required. As suggested in [211] this model could be combined with the concept of dynamic topic models to model both how the topics and the underlying dimensionality of the vocabulary changes. This could be used to identify changes in the variety of activities a person performs.

The results of applying dynamic topic models to activity data to identify changes in the structure of daily routines have limitations, as discussed in section 5.4. To expand on this work the next step would be to run a larger experiment on more complex simulated datasets with multiple parameters changing in any permutation. This would demonstrate the ability of DTMs to detect changes in data that are more similar to a real-world dataset. If DTMs can successfully identify changes in a more complex simulated dataset then they can be tested on real-world long term data with known changes. Furthermore, valid criteria for detecting significant changes related to health and well-being need to be established in collaboration with relevant experts.

Another important area of research for applying DTMs to activity data is to establish the optimal vocabulary and length of small time slices to discover changes at the activity level. These decisions should be made in conjunction with expert clinical opinions in order to be able to identify changes that are significant in terms of healthcare. Different models may be required for different situations if changes are occurring at varying levels of granularity.

In addition to DTMs there are several other extensions to LDA, as described in section 2.3.2. In particular, Hierarchical Dirichlet Processes (HDP) could be applied to activity data to discover routines without specifying the number in advance as this is determined directly by the model. HDP also offers the advantage of allowing new topics or routines to be evoked by previously unseen data. Therefore if new activities occur that do not fit an existing routine, a new one can be defined. This model can then be further extended to create a model known as dynamic Hierarchical Dirichlet Processes that models how topic proportions change over time. This could be used to identify changes over time at a different level i.e. the variation in the probability of routines over time.

As well as overcoming the limitation of needing to define the number of routines for LDA in advance, Seiter et al. [235] propose an approach that addresses the limitation of time-invariant segments, which include failure to handle transitions, variations in activity duration, and short activities accurately. This is a hierarchical topic model approach that includes automatic segmentation of raw sensor data into context words in an unsupervised way, avoiding the need for annotations. It allows segments to vary in size by performing segmentation dynamically based on the data. The hierarchical model uses a segmentation prior, that considers semantic and temporal features of context words, and distance dependent Chinese restaurant process (ddCRP) to group segments and the Chinese restaurant process (CRP) to discover activities. Future work could build on this approach to address these limitations in DTMs for discovering changes in routines over time.

8.2.2 Future work for detecting eating event types

There is a clear opportunity to extend the work presented on detecting different types of eating events. The discovered combinations of food groups can be used, with or without subjective labels, as a basis for further analysis to address a wide range of nutrition research questions. For example, patterns in eating behaviours of individuals can be explored further to determine if there are associations with specific health and well being outcomes. This could be applied to the data from the four and seven day food diaries in the datasets utilised in this research or longer term data to find more detailed patterns.

At a population level, characteristic patterns of specific groups, such as children, could be investigated. For example, visualising the probability of different eating events over time might reveal that main meal type events are more likely between 5 and 7pm than for a group of adults aged 19 - 64. The discovered eating event types could also be linked back to other variables of interest recorded in the datasets. For example, investigating correlations between a discovered eating event type and the energy or nutrient content of the eating events for which this type is the most likely.

The effect of breakfast on the remaining eating event types throughout the rest of the day was considered in section 7.1.1.2. It was observed that people consuming ‘breakfast’ between 6 and 10am were more likely to have a second breakfast. The reason for this could be that the first event labelled as breakfast was just a cup of tea or coffee and a more substantial breakfast or second drink was consumed later. Investigating this further by considering the energy of the eating events or removing drink only events could reveal whether this hypothesis is correct.

The results highlighted that eating events containing all of the individual items for a homemade dessert e.g. a cake, are often assigned to an incorrect eating event type. This is because the food groups of the individual items i.e. eggs, flour, milk, butter are different in nature to that for the final product. On the other hand, for some main meals e.g. beef stew, that are not split into their component parts, the incorrect label is assigned due to the lack of detail. Eating events with only one or two items and drink only events are also often mislabelled. An interesting avenue for further work would be to investigate whether rule-based preprocessing of the data could resolve these limitations.

Various extensions to LDA could also be applied to nutrition data. In particular, HDP could be investigated so that the number of eating events is automatically determined by the model. However, this may not improve the results as the best fit to the data and the semantic coherence of the results do not always align. A variation on author topic models could also be considered as a solution for including more information about eating events. For example, the weight, energy and time of the eating event, that were identified as key properties could be the equivalent of the author of a document. This means that the link between these variables and the food groups are explicitly modelled. However, the original author topic model assumes there is one author for a document containing multiple topics. Therefore the model would have to be modified to reflect the nature of the nutrition application.

Appendices

Appendix A

Activity Vocabularies

This appendix provides reference lists of the vocabularies of activities used for different models that were not included in the main body of the thesis.

Table A.1: Custom vocabulary from ADL Logger App for participant two. Each activity in the vocabulary has an ID and a description and is associated with a routine (1-9).

Activity		**Routine**	
ID	Description	ID	Description
1	Showering	1	Wake Up
2	Personal hygiene	1	Wake Up
3	Eating breakfast	1	Wake Up
4	Eating snack	8	Multi
5	Working at desk	3	Work
6	Attending meeting	3	Work
7	Bus	2	Commute
8	Train	2	Commute
9	Dressing	1	Wake Up
10	Yoga	6	Exercise
11	Phone call	8	Multi
12	Using toilet	8	Multi
13	Walking	8	Multi
14	Watching television	7	Relaxation
15	Sleeping	9	Other
16	Walking	8	Multi
17	Preparing drink	8	Multi
18	Eating lunch	4	Lunch

Activity		Routine	
ID	Description	ID	Description
19	Eating dinner	5	Dinner
20	Preparing dinner	5	Dinner
21	Shopping	9	Other
22	Waiting	8	Multi
23	Drinking	8	Multi
24	Tidying up	9	Other
25	Preparing breakfast	1	Wake Up
26	Doing chores	9	Other
27	Using computer	7	Relaxation
28	Reading	7	Relaxation
29	Relaxing	7	Relaxation
30	Preparing lunch	4	Lunch
31	Personal hygiene	8	Multi
32	Sitting talking	8	Multi
33	Car	2	Commute
34	Golf	6	Exercise
35	Ironing	9	Other
36	Doing other work	3	Work
37	Preparing food	8	Multi

Table A.2: List of activities at the low level of abstraction defined by the Home Sensor Simulator. Note - the typing mistakes are kept as these are the labels directly generated by the simulation.

Activity		Activity	
ID	Description	ID	Description
1	brush_teeth	39	go_kitchen_shelf
2	change_clothes	40	go_kitchen_sink
3	cook_and_eat	41	go_micro
4	do_exercise	42	go_outside
5	do_the_dishes	43	go_oven
6	do_walk	44	go_shoe_shelf
7	do_watch_tv	45	go_tv
8	dress_down_outdoor	46	go_tv_chair
9	dress_up_outdoor	47	go_wardrobe
10	drink	48	go_wc

Activity		Activity	
ID	Description	ID	Description
11	drink_water	49	have_bath
12	eat_cold	50	interact_with_man
13	eat_cold_meal	51	nonprepared_sleep
14	eat_warm	52	pack_food
15	eat_warm_meal	53	pack_goods
16	exercise	54	plate_to_sink
17	finish_walk	55	prepared_sleep
18	get_bread	56	put_meal_to_fridge
19	get_clothes	57	put_plate_to_sink
20	get_cold_ingredients	58	rest
21	get_cold_warm_food	59	rest_in_chair
22	get_food	60	shop
23	get_food_from_fridge	61	sleep_in_bed
24	get_glass	62	someone_at_entrace
25	get_ingeredients_from_fridge	63	switch_computer_off
26	get_ingeredients_from_shelf	64	switch_computer_on
27	get_water	65	switch_tv_off
28	go_bathroom_sink	66	switch_tv_on
29	go_bathtub	67	use_computer
30	go_bed	68	use_micro
31	go_chair	69	use_oven
32	go_computer	70	use_the_computer
33	go_computer_chair	71	walk_outside
34	go_dining_chair	72	wash_dishes
35	go_dining_table	73	wash_hands
36	go_entrace	74	wc
37	go_exercise_place	75	wc_do
38	go_fridge	76	wc_flush

Table A.3: List of activities at the higher level of abstraction defined by the Home Sensor Simulator. Note - the typing mistakes are kept as these are the labels directly generated by the simulation.

Activity	
ID	Description
1	cook_and_eat
2	do_the_dishes
3	do_walk
4	drink
5	eat_cold
6	eat_warm
7	exercise
8	prepared_sleep
9	rest
10	shop
11	someone_at_entrace
12	use_computer
13	watch_tv
14	wc

Appendix B

Food Group Lists

This appendix provides the lists of food groups from each of the datasets used for reference. The lists of meal, snack and drink foods used for the rule-based labelling method published as supplementary material in [202] is also given for reference.

Table B.1: Food [182] and day [227] level food group lists from NDNS RP dataset

NDNS RP Food Level Food Groups		**NDNS RP Day Level Food Groups**	
ID	Description	ID	Description
1	pasta rice and other cereals	0	bacon and ham
2	white bread	1	beef veal and dishes
3	wholemeal bread	2	beer lager cider perry
4	other bread	3	biscuits
5	high fibre breakfast cereals	4	brown granary and wheat germ bread
6	other breakfast cereals	5	buns cakes pastries fruit pies
7	biscuits	6	burgers and kebabs
8	buns cakes pastries & fruit pies	7	butter
9	puddings	8	cheese
10	whole milk	9	chicken and turkey dishes
11	semi skimmed milk	10	chips fried roast potatoes and potato products
12	skimmed milk	11	chocolate confectionery
13	other milk and cream	12	coated chicken
14	cheese	13	commercial toddlers foods and drinks
15	yogurt fromage frais and dairy desserts	14	crisps and savoury snacks

NDNS RP Food Level Food Groups		NDNS RP Day Level Food Groups	
ID	Description	ID	Description
16	eggs and egg dishes	15	eggs and egg dishes
17	butter	16	fruit
18	pufa margarine & oils	17	fruit juice
19	low fat spread	18	high fibre breakfast cereals
20	other margarine fats and oils	19	ice cream
21	reduced fat spread	20	lamb and dishes
22	bacon and ham	21	liver dishes
23	beef veal and dishes	22	low fat spread
24	lamb and dishes	23	meat pies and pastries
25	pork and dishes	24	nuts and seeds
26	coated chicken	25	oily fish
27	chicken and turkey dishes	26	one percent milk
28	liver & dishes	27	other bread
29	burgers and kebabs	28	other breakfast cereals
30	sausages	29	other margarine fats and oils
31	meat pies and pastries	30	other meat and meat products
32	other meat and meat products	31	other milk and cream
33	white fish coated or fried	32	other potatoes potato salads dishes
34	other white fish shellfish & fish dishes	33	other white fish shell fish fish dishes
35	oily fish	34	pasta rice and other cereals
36	salad and other raw vegetables	35	pork and dishes
37	vegetables not raw	36	puddings
38	chips fried & roast potatoes and potato products	37	pufa margarine oils
39	other potatoes potato salads & dishes	38	reduced fat spread
40	fruit	39	salad and other raw vegetables
41	sugars preserves and sweet spreads	40	sausages
42	crisps and savoury snacks	41	semi skimmed milk
43	sugar confectionery	42	skimmed milk
44	chocolate confectionery	43	soft drinks low calorie
45	fruit juice	44	soft drinks not low calorie
47	spirits and liqueurs	45	spirits and liqueurs
48	wine	46	sugar confectionery

NDNS RP Food Level Food Groups		NDNS RP Day Level Food Groups	
ID	Description	ID	Description
49	beer lager cider & perry	47	sugars preserves and sweet spreads
50	miscellaneous	48	tea coffee and water
51	tea coffee and water	49	vegetables not raw
52	commercial toddlers foods and drinks	50	white bread
53	ice cream	51	white fish coated or fried
54	dietary supplements	52	wholemeal bread
55	artificial sweeteners	53	whole milk
56	nuts and seeds	54	wine
57	soft drinks not low calorie	55	yogurt fromage frais and dairy desserts
58	soft drinks low calorie	56	dry weight beverages
59	brown granary and wheatgerm bread	57	low fat spread not polyunsaturated
60	1% fat milk	58	polyunsaturated low fat spread
61	smoothies 100% fruit and/or juice	59	reduced fat spread not polyunsaturated
		60	reduced fat spread polyunsaturated
		61	savoury sauces pickles gravies condiments
		62	soup homemade and retail
		63	cheddar cheese
		64	cottage cheese
		65	other cheese
		66	smoothies 100 fruit and/or juice

Table B.2: Food group lists for NDNS 2000 [233] and NANS [183] datasets

NDNS 2000 Food Groups		NANS Food Groups	
ID	**Description**	**ID**	**Description**
1	Pasta, rice and other cereals	1	Rice & pasta, flours, grains & starch
2	White bread	2	Savouries
3	Wholemeal bread	3	White sliced bread & rolls
4	Other breads	4	Wholemeal & brown bread & rolls
5	High fibre breakfast cereals	5	Other breads
6	Other breakfast cereals	6	RTEBC
7	Biscuits	7	Other breakfast cereals
8	Buns, cakes, pastries and fruit pies	8	Biscuits including crackers
9	Puddings	9	Cakes, pastries & buns
10	Whole milk	10	Whole milk
11	Semi-skimmed milk	11	Low fat, skimmed & fortified milks
12	Skimmed milk	12	Other milks & milk based beverages
13	Other milk and cream	13	Creams
14	Cheese	14	Cheeses
15	Yogurt and other dairy desserts	15	Yoghurts
16	Eggs and egg dishes	16	Ice-creams
17	Butter	17	Desserts
18	PUFA margarine and oils	18	Rice puddings & custard
19	Low fat spread	19	Eggs & egg dishes
20	Margarine and other fats, not PUFA	20	Butter (over 80% fat)
21	Reduced fat spread	21	Low fat spreads (under 40% fat)
22	Bacon and ham	22	Other fat spreads (40-80% fat)
23	Beef, veal and dishes	23	Oils (not including those used in recipes)
24	Lamb and dishes	24	Hard cooking fats
25	Pork and dishes	25	Potatoes (boiled/baked/mashed)
26	Coated chicken and turkey	26	Processed & homemade potato products

NDNS 2000 Food Groups		NANS Food Groups	
ID	Description	ID	Description
27	Chicken and turkey dishes	27	Chipped, fried & roasted potatoes
28	Liver, liver products and dishes	28	Vegetable & pulse dishes
29	Burgers and kebabs	29	Peas, beans & lentils
30	Sausages	30	Green vegetables
31	Meat pies and pastries	31	Carrots
32	Other meat and meat products	32	Salad vegetable
33	Fried white fish	33	Other vegetables
34	Other white fish, shellfish and fish dishes	34	Tinned or jarred vegetables
35	Oily fish	35	Fruit juices & smoothies
36	Salad and other raw vegetables	36	Bananas
37	Vegetables, not raw	37	Other fruits
38	Chips, fried and roast potatoes and products	38	Citrus fruits
39	Other potatoes, potato salads and dishes	39	Tinned fruits
40	Fruit	40	Nuts & seeds, herbs & spices
41	Sugars, preserves and sweet spreads	41	Fish & fish products
42	Crisps and savoury snacks	42	Fish dishes
43	Sugar confectionery	43	Bacon & ham
44	Chocolate confectionery	44	Beef & Veal
45	Fruit juice	45	Lamb
47	Spirits and liqueurs	46	Pork
48	Wine	47	Chicken, turkey & game
49	Beer, lager, cider and perry	48	Offal & offal dishes
50	Miscellaneous	49	Beef & veal dishes
51	Tea, coffee and water	50	Lamb, pork & bacon dishes
52	Commercial toddlers foods and drinks	51	Poultry & game dishes
53	Ice cream	52	Burgers
54	Dietary supplements	53	Sausages
55	Artificial sweetners	54	Meat pies & pastries
56	Nuts and Seeds	55	Meat products
57	Soft drinks, not low calorie	56	Alcoholic beverages

NDNS 2000 Food Groups		NANS Food Groups	
ID	Description	ID	Description
58	Soft drinks, low calorie	57	Sugars, syrups, preserves & sweetners
		58	Chocolate confectionary
		59	Non-chocolate confectionary
		60	Savoury snacks
		61	Soups, sauces & miscellaneous foods
		62	Nutritional supplements
		63	Teas
		64	Coffees
		65	Other beverages
		66	Carbonated beverages
		67	Diet carbonated beverages
		68	Squashes, cordials & fruit juice drinks

Table B.3: Meal, snack and drink lists for rule based labels method [202]

Meal Food	Snack Food	Drink
Bacon & ham	Apples pears	Alco-pops
Baked beans	Bananas	Beers
Beef, veal	Biscuits	Other beverage
Burgers kebab	Block marge	Bottled water
Carrot	Buns cakes pastries	Cider perry
Chicken & turkey	Chocolate con	Coffee
Coated chicken	Cottage cheese	Common toddler drinks
Common toddler food	Cream	Fortified wine
Cooked tomatoes	Fromage frais	Fruit juice
Egg dishes	Fruit in juice	Herbal tea
Eggs	Fruit in syrup	Liqueurs
Fried white fish	Fruit pies	Low alc beers
Green beans	Ice cream	Low alc cider perry
Lamb	Low fat spread	Low alc wine
Leafy green	Milk puds	Other milk
Liver	Nuts & seeds	Semi-skimmed
Meat pies etc	Oils & fats not pufa	Skimmed milk
Non fried potato products	Oranges	Soft drink fizzy diet
Oily fish	Other dairy dessert	Soft drink fizzy non-diet
Other white fish	Other cheese	Soft drink squash diet
Other breakfast cereals	Other fruit	Soft drink squash non-diet
Other bread	Other puds	Soft drink still diet
Other cereal	Other salad	Soft drink still non-diet
Other fried pots	Other sugars	Spirits
Other meat	Preserves	Tap water
Other potato dishes	PUFA low fat spread	Tea
Other veg	PUFA marge	Whole milk
Pasta	PUFA oils	Wine
Peas	PUFA reduced fat spread	
Pizza	Raw carrots	
Pork	Raw tomatoes	
Potato chips	Reduced fat spread	
Rice	Savoury sauces	
Sausages	Savoury snacks	
Shellfish	Soft marge not PUFA	
Softgrain bread	Sponge puddings	

Meal Food	**Snack Food**	**Drink**
Soups	Sugar	
Vegetable dishes	Sugar confectionery	
High fibre breakfast cereal	Sweeteners	
White bread	Yogurt	
Wholemeal bread	Butter	

Appendix C

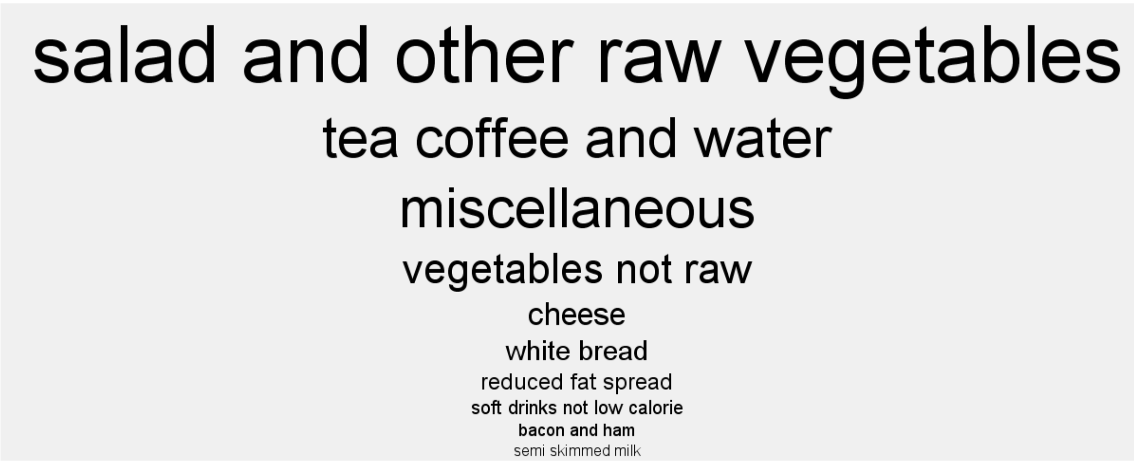
Detected Eating Event Types

C.1 NDNS RP 10 Type Model

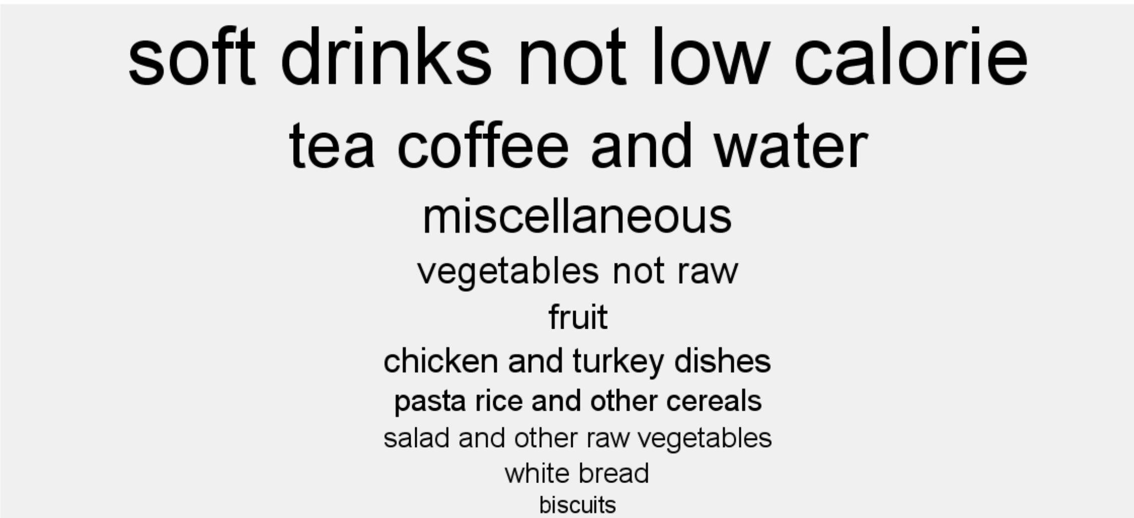
(a) Topic 1 : Main meal

Figure C.1: All 10 eating event types and labels for NDNS RP 10 Type model

(b) Topic 2 : Breakfast



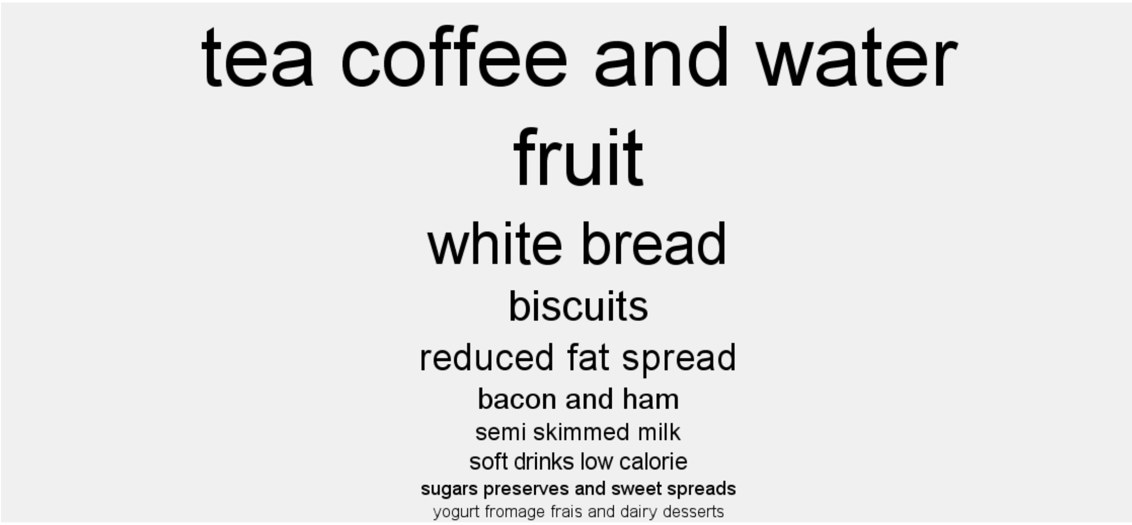
(c) Topic 3 : Light meal



(d) Topic 4 : Main meal

Figure C.1: NDNS RP 10 Type model

(e) Topic 5 : Snack



(f) Topic 6 : Light meal

(g) Topic 7 : Snack

Figure C.1: NDNS RP 10 Type model

(h) Topic 8 : Snack

(i) Topic 9 : Breakfast

(j) Topic 10 : Main meal

Figure C.1: NDNS RP 10 Type model

C.2 NDNS RP 15 Type Model

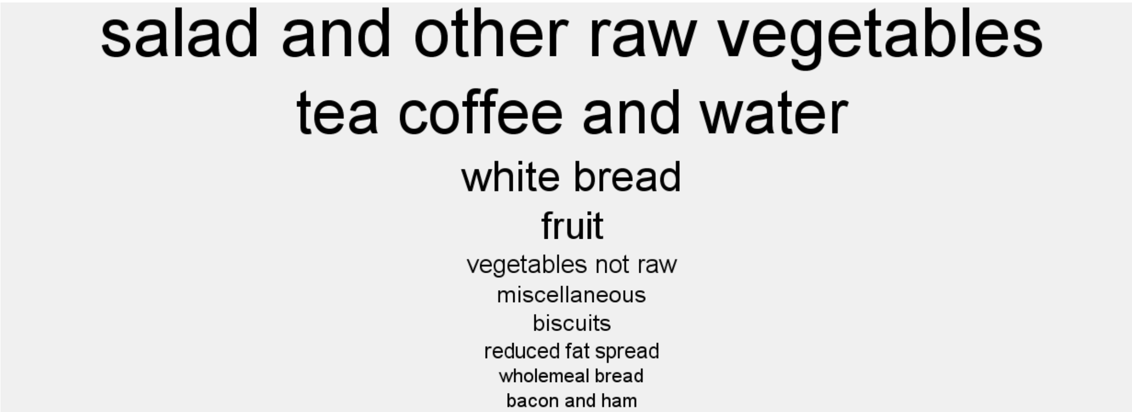
(a) Topic 1 : Snacks

(b) Topic 2 : Breakfast

(c) Topic 3 : Drinks

Figure C.2: All 15 eating event types and labels NDNS RP 15 Type model

(d) Topic 4 : Mixed



(e) Topic 5 : Lunch/sandwiches

(f) Topic 6 : Main meal

Figure C.2: NDNS RP 15 Type model

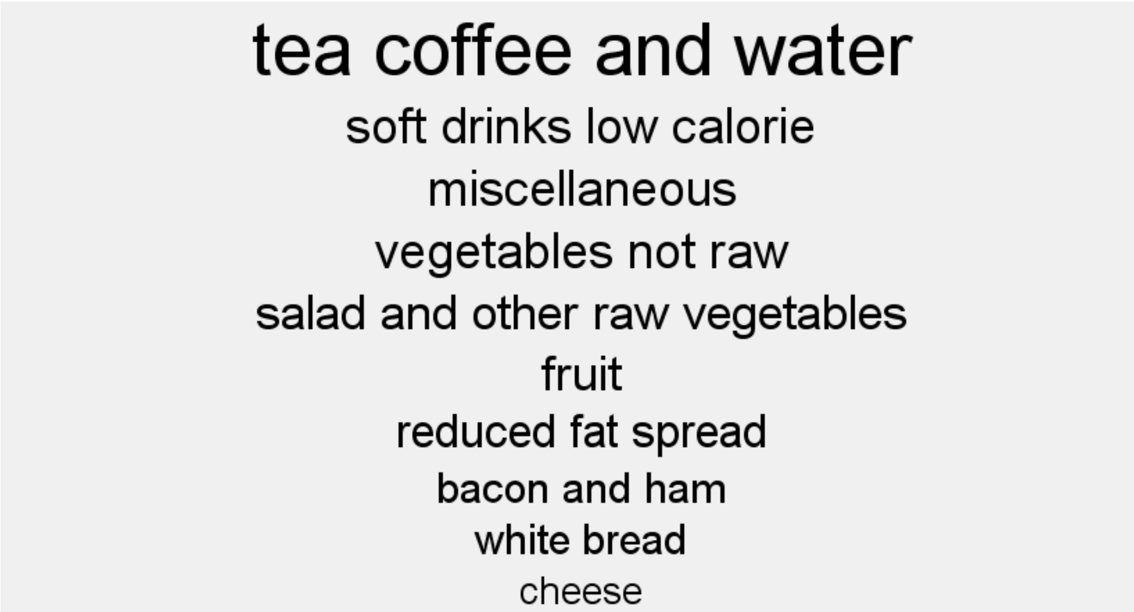
(g) Topic 7 : Snacks

(h) Topic 8 : Light meal

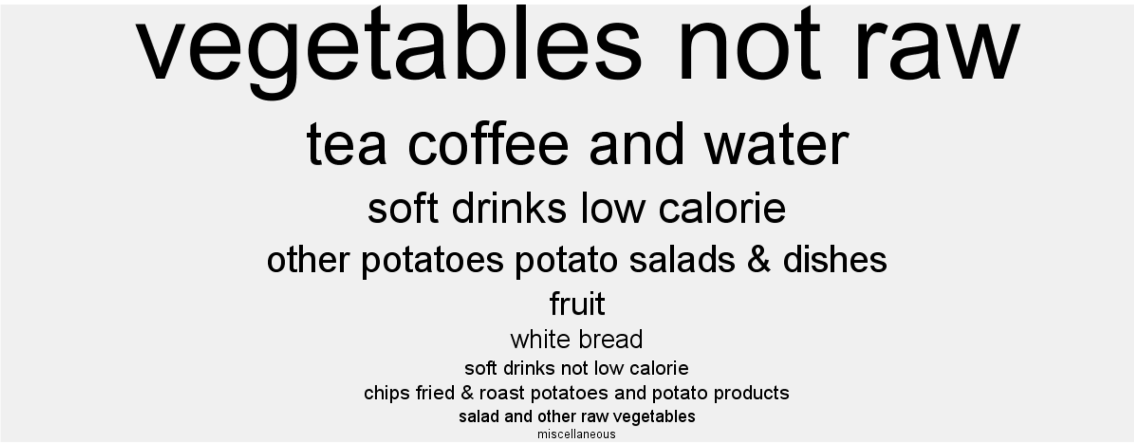
(i) Topic 9 : Drinks and snacks

Figure C.2: NDNS RP 15 Type model

(j) Topic 10 : Snacks

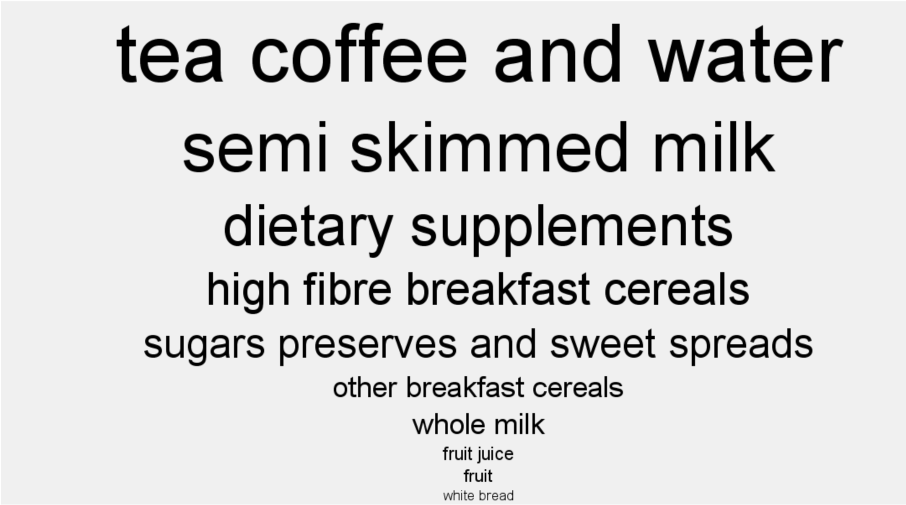


(k) Topic 11 : Lunch/sandwiches

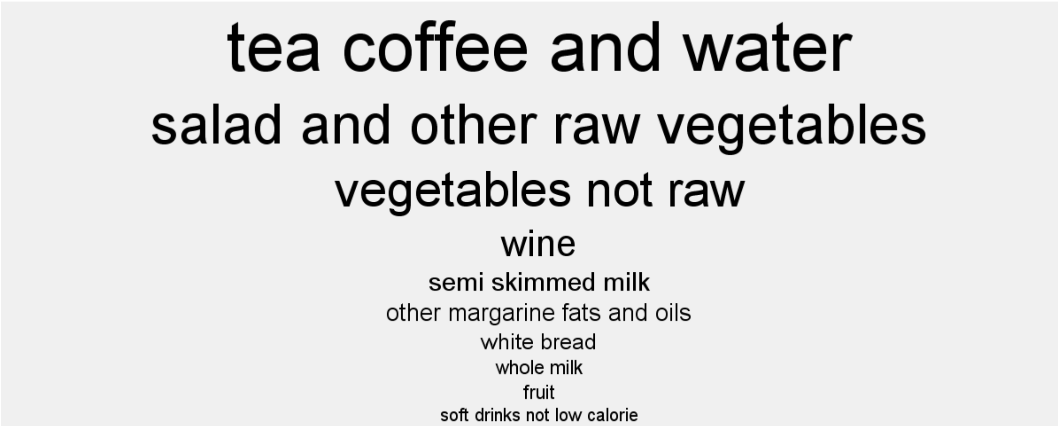


(l) Topic 12 : Main meal

Figure C.2: NDNS RP 15 Type model



(m) Topic 13 : Breakfast

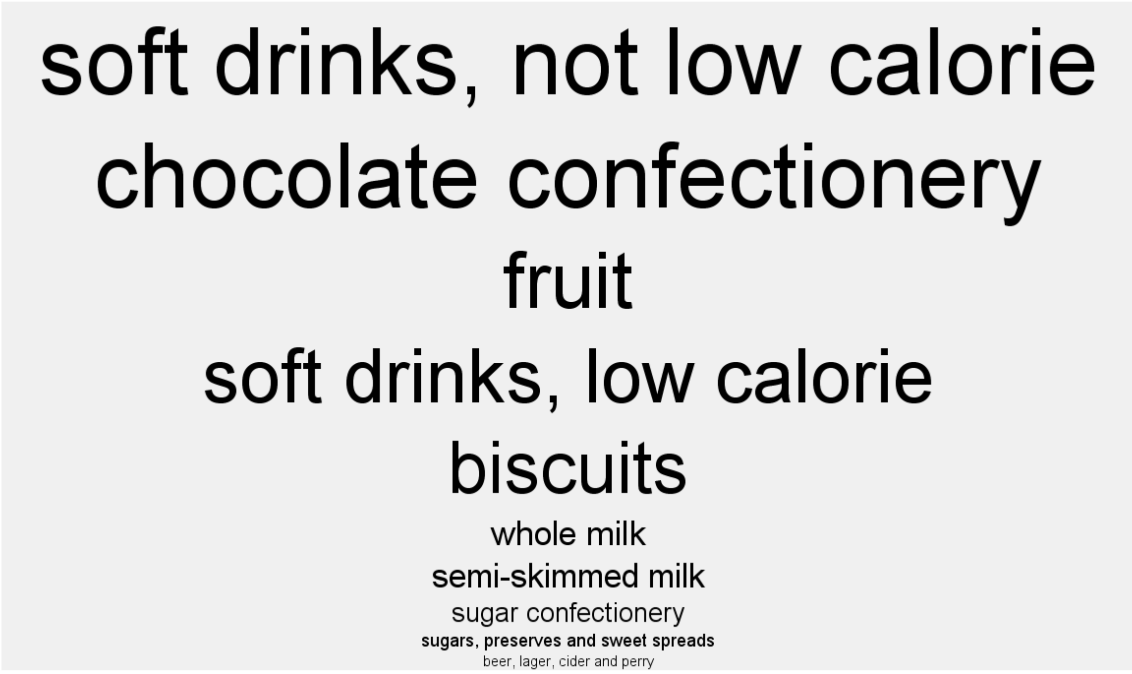


(n) Topic 14 : Main meal

(o) Topic 15 : Lunch/sandwiches

Figure C.2: NDNS RP 15 Type model

C.3 NDNS 2000 5 Type Model



(a) Topic 1 : Snacks

(b) Topic 2 : Drinks

Figure C.3: All 5 eating event types and labels NDNS 2000 5 Type model

(c) Topic 3 : Main meal

(d) Topic 4 : Breakfast

(e) Topic 5 : Drinks

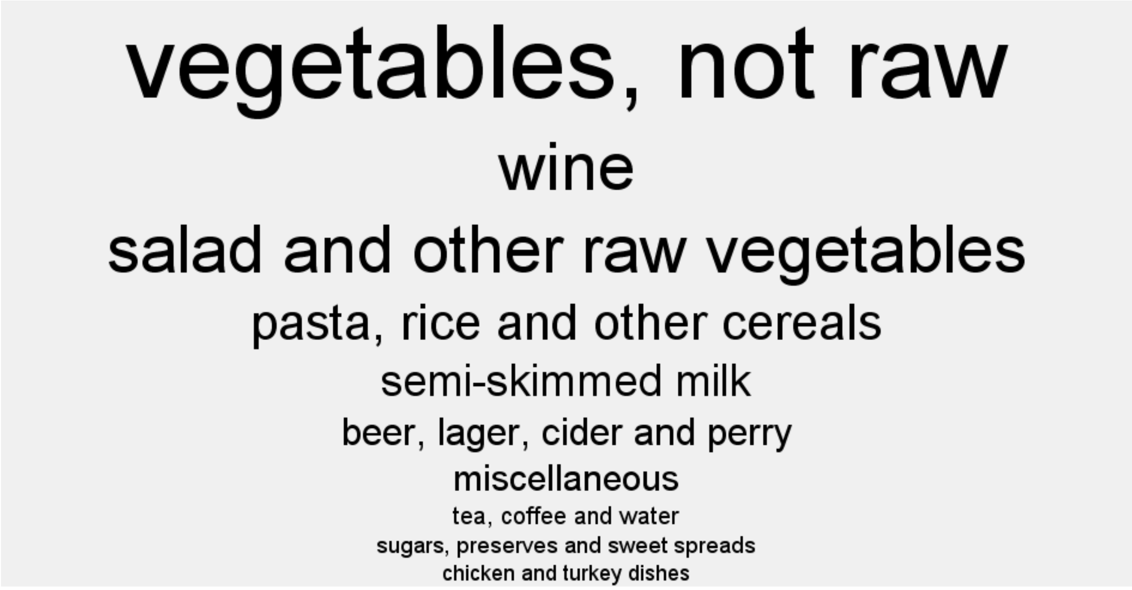
Figure C.3: NDNS 2000 5 Type model

C.4 NDNS 2000 10 Type Model

(a) Topic 1 : Main meal

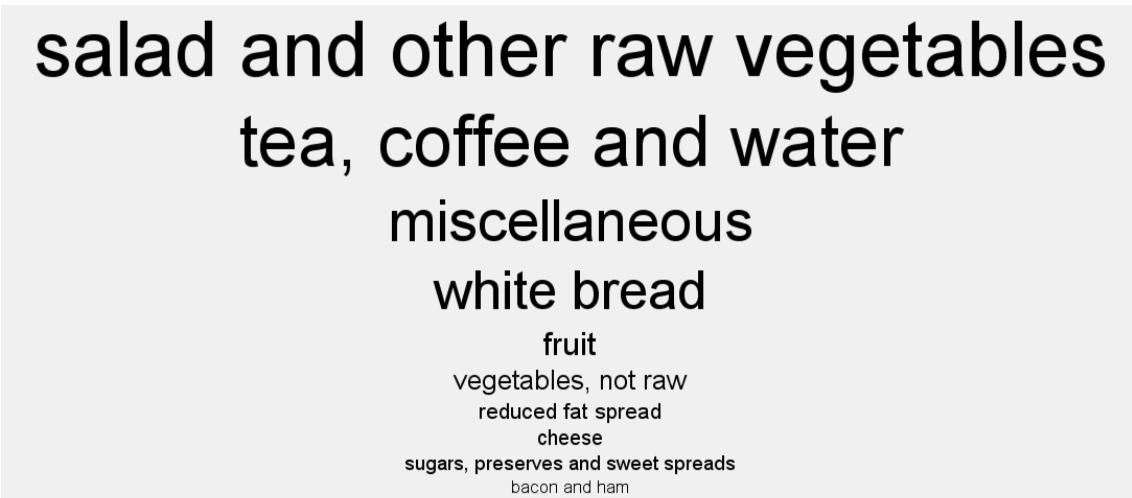
(b) Topic 2 : Breakfast

Figure C.4: All 10 eating event types and labels NDNS 2000 10 Type model



(c) Topic 3 : Main meal

(d) Topic 4 : Snacks



(e) Topic 5 : Lunch/sandwiches

Figure C.4: NDNS 2000 10 Type model

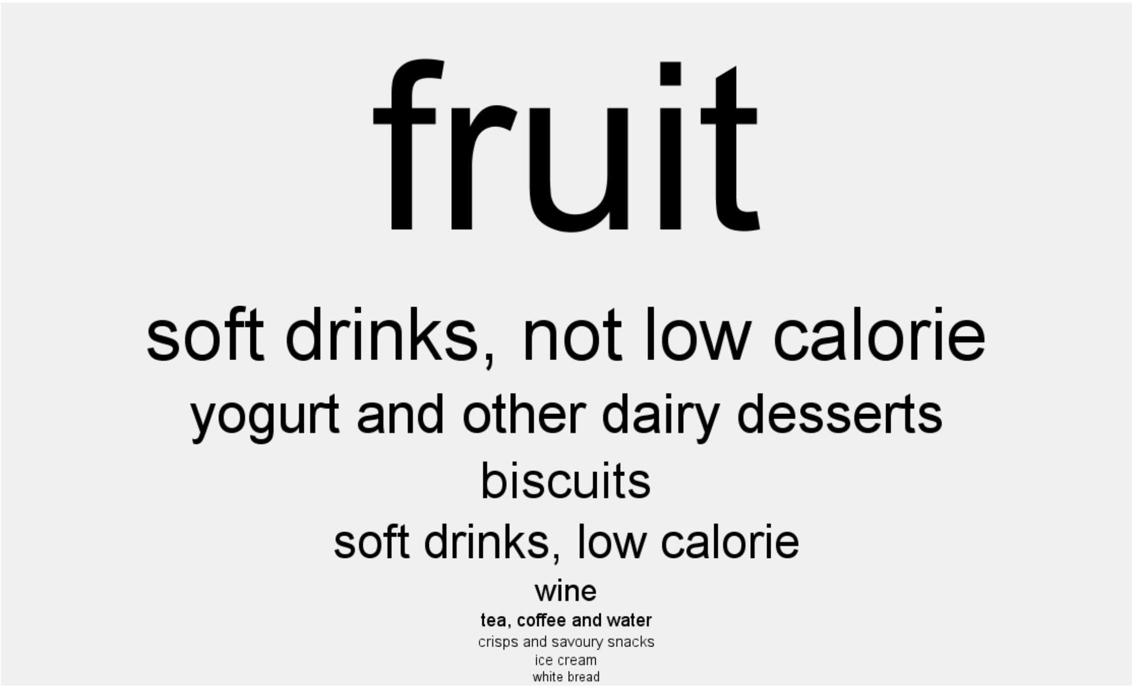
(f) Topic 6 : Light meal

(g) Topic 7 : Lunch/sandwiches

(h) Topic 8 : Drinks

Figure C.4: NDNS 2000 10 Type model

(i) Topic 9 : Drinks



(j) Topic 10 : Snacks

Figure C.4: NDNS 2000 10 Type model

C.5 NDNS 2000 15 Type Model

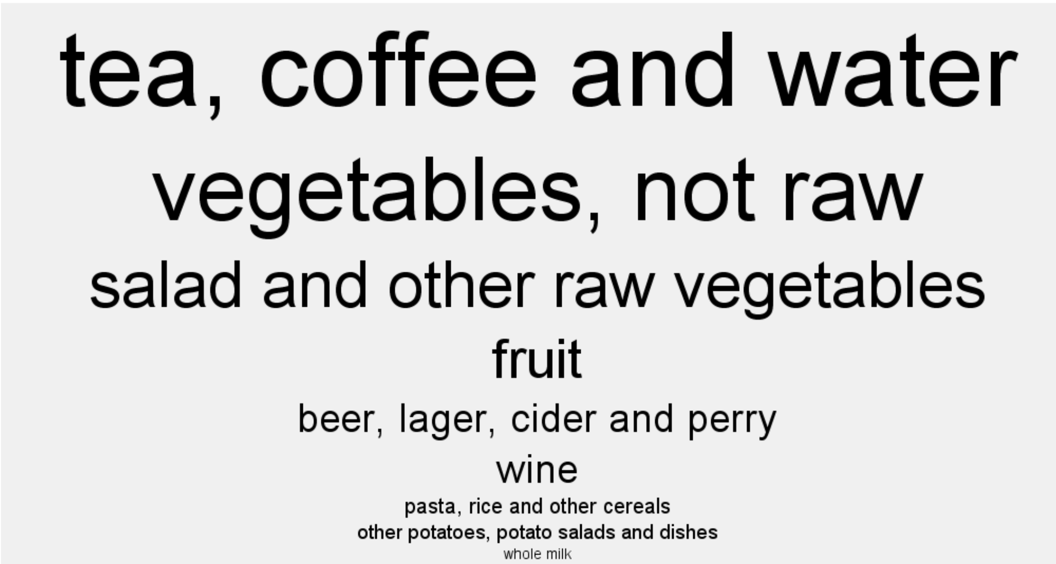
(a) Topic 1 : Snacks

(b) Topic 2 : Breakfast

(c) Topic 3 : Snacks

Figure C.5: All 15 eating event types and labels NDNS 2000 15 Type model

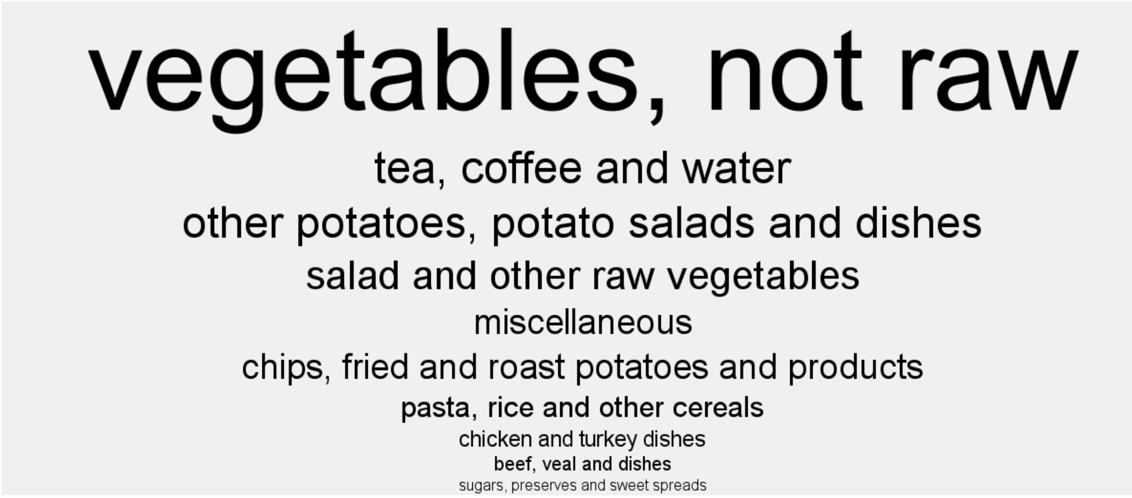
(d) Topic 4 : Light meal



(e) Topic 5 : Light meal

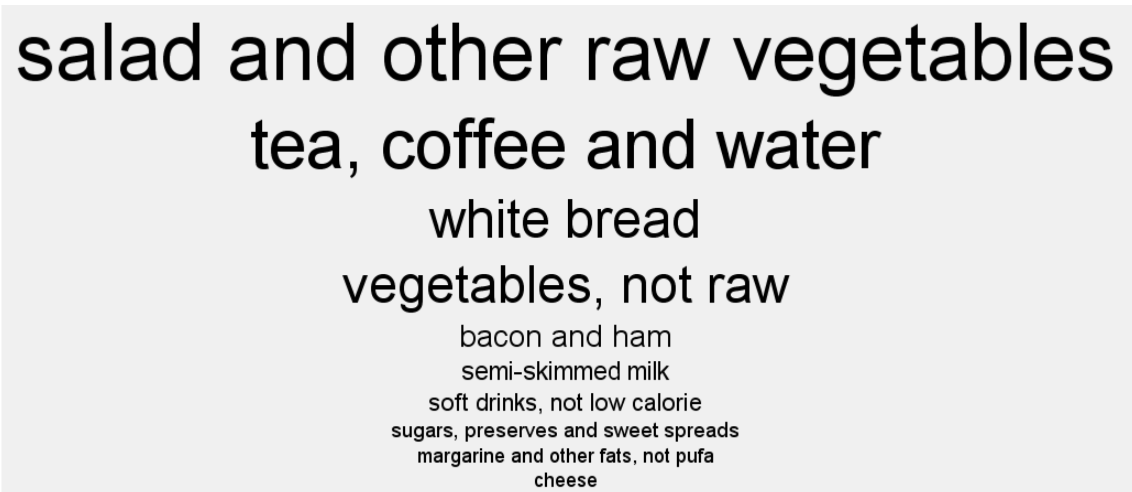
(f) Topic 6 : Snacks

Figure C.5: NDNS 2000 15 Type model



(g) Topic 7 : Main meal

(h) Topic 8 : Lunch/sandwiches



(i) Topic 9 : Lunch/sandwiches

Figure C.5: NDNS 2000 15 Type model

(j) Topic 10 : Drinks

(k) Topic 11 : Snacks

(l) Topic 12 : Breakfast

Figure C.5: NDNS 2000 15 Type model

(m) Topic 13 : Breakfast

(n) Topic 14 : Drinks

(o) Topic 15 : Lunch/sandwiches

Figure C.5: NDNS 2000 15 Type model

C.6 NANS 5 Type Model

(a) Topic 1 : Breakfast

(b) Topic 2 : Light meal

Figure C.6: All 5 eating event types and labels NANS 5 Type model

(c) Topic 3 : Drinks and snacks

(d) Topic 4 : Main meal

(e) Topic 5 : Drinks and snacks

Figure C.6: NANS 5 Type model

C.7 NANS 10 Type Model

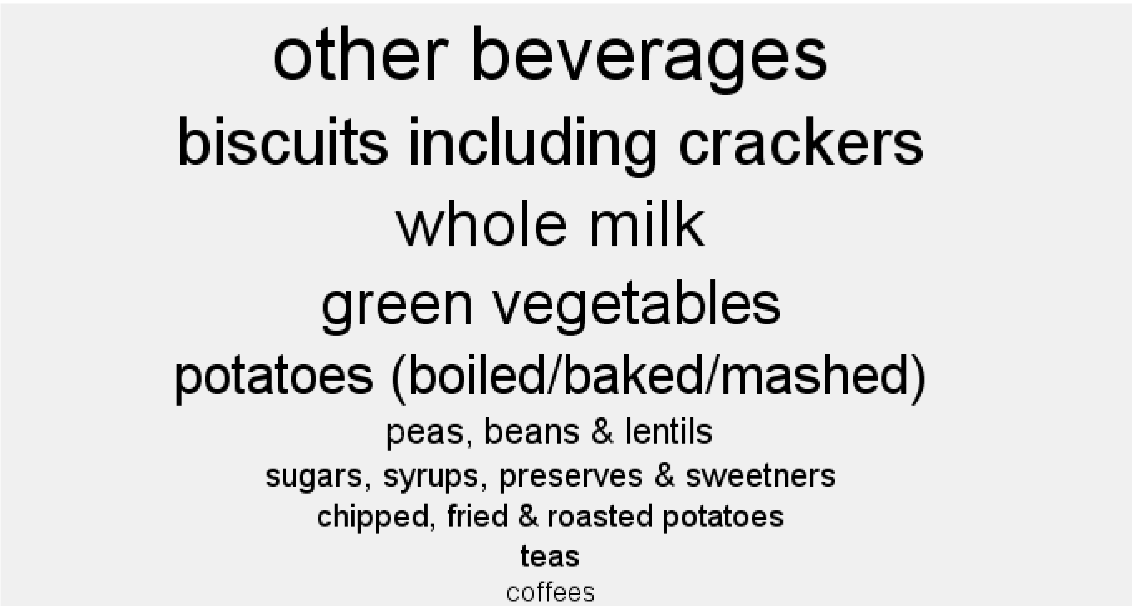
(a) Topic 1 : Drinks and snacks

(b) Topic 2 : Light meal

(c) Topic 3 : Main meal

Figure C.7: All 10 eating event types and labels NANS 10 Type model

(d) Topic 4 : Breakfast



(e) Topic 5 : Main meal

Figure C.7: NANS 10 Type model

(f) Topic 6 : Drinks and snacks

(g) Topic 7 : Lunch/sandwiches

(h) Topic 8 : Lunch/sandwiches

Figure C.7: NANS 10 Type model

(i) Topic 9 : Breakfast

(j) Topic 10 : Main meal

Figure C.7: NANS 10 Type model

C.8 NANS 15 Type Model

(a) Topic 1 : Lunch/sandwiches

(b) Topic 2 : Lunch/sandwiches

(c) Topic 3 : Lunch/sandwiches

Figure C.8: All 15 eating event types and labels NANS 15 Type model

(d) Topic 4 : Drinks and snacks

(e) Topic 5 : Main meal

(f) Topic 6 : Breakfast

Figure C.8: NANS 15 Type model

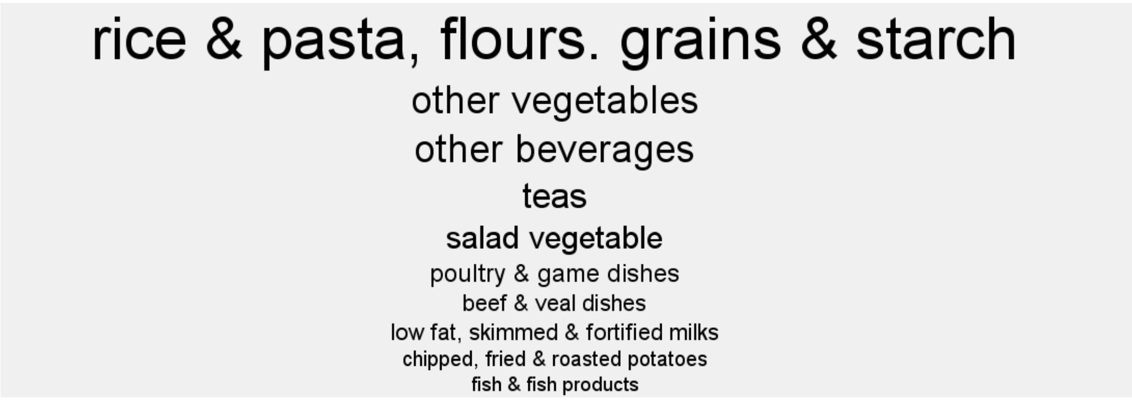
(g) Topic 7 : Main meal

(h) Topic 8 : Lunch/sandwiches

(i) Topic 9 : Lunch/sandwiches

Figure C.8: NANS 15 Type model

(j) Topic 10 : Drinks (alcoholic)



(k) Topic 11 : Main meal

(l) Topic 12 : Breakfast

Figure C.8: NANS 15 Type model

(m) Topic 13 : Light meal

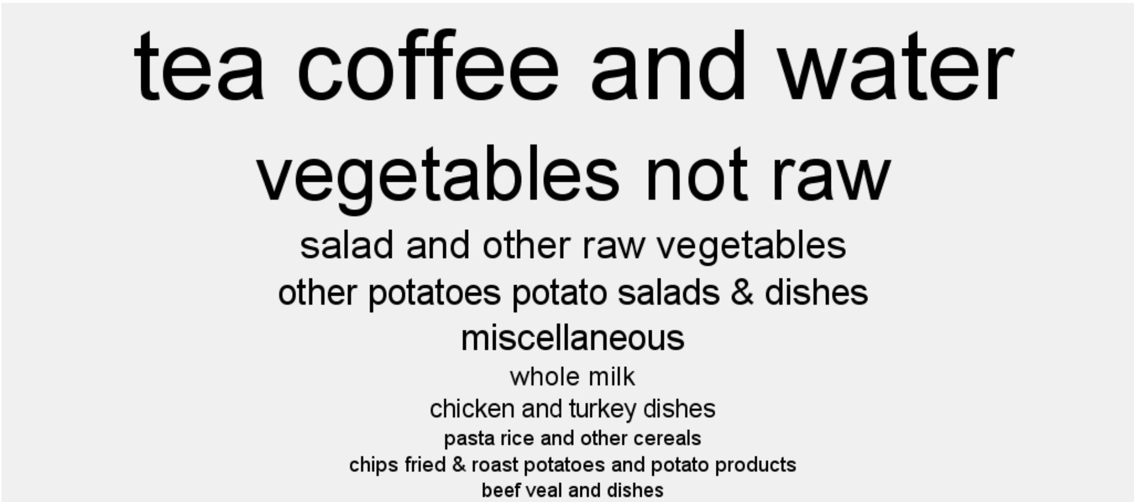
(n) Topic 14 : Main meal

(o) Topic 15 : Light meal

Figure C.8: NANS 15 Type model

C.9 NANS with NDNS food groups 5 Type Model

(a) Topic 1 : Breakfast

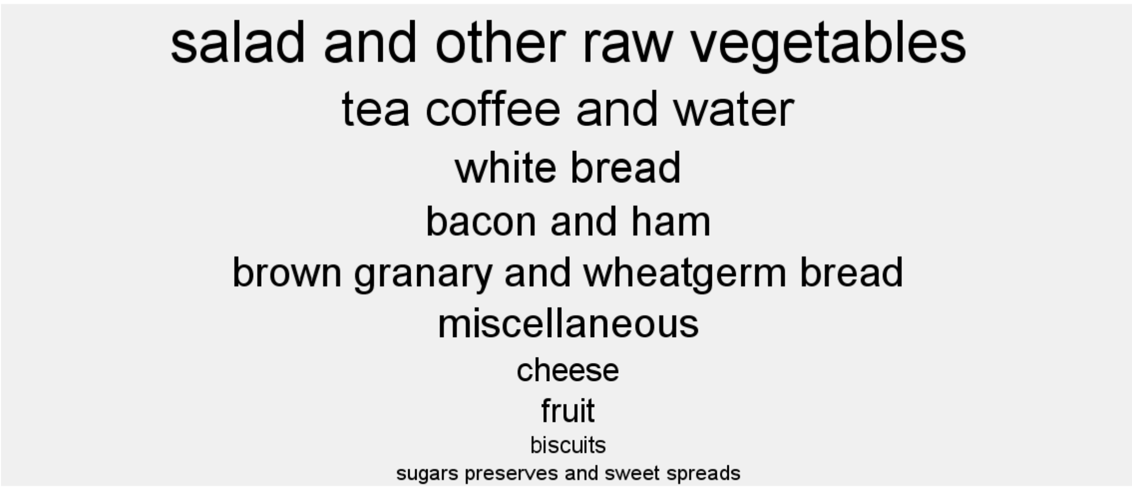


(b) Topic 2 : Main meal

Figure C.9: All 5 eating event types and labels NANS with NDNS RP food groups 5 Type model

(c) Topic 3 : Drinks and snacks

(d) Topic 4 : Drinks and snacks



(e) Topic 5 : Lunch/sandwiches

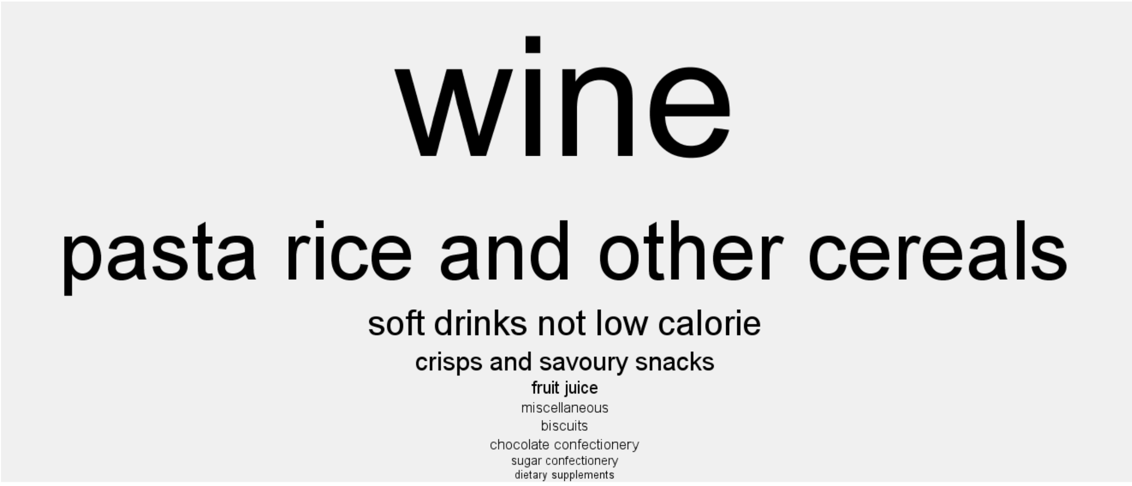
Figure C.9: NANS with NDNS RP food groups 5 Type model

C.10 NANS with NDNS food groups 10 Type Model

(a) Topic 1 : Breakfast

(b) Topic 2 : Lunch/sandwiches

Figure C.10: All 10 eating event types and labels NANS with NDNS RP food groups 10 Type model



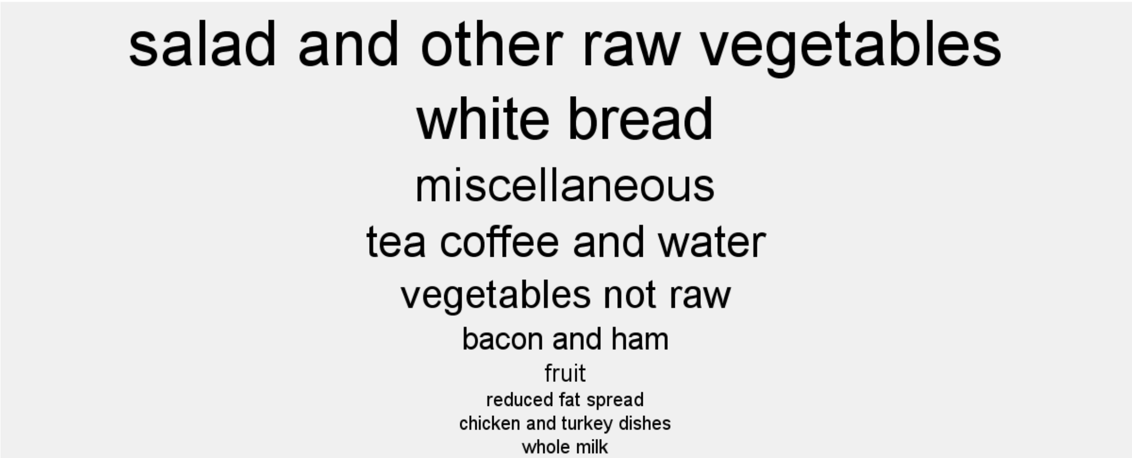
(c) Topic 3 : Drinks and snacks

(d) Topic 4 : Main meal

(e) Topic 5 : Main meal

Figure C.10: NANS with NDNS RP food groups 10 Type model

(f) Topic 6 : Breakfast

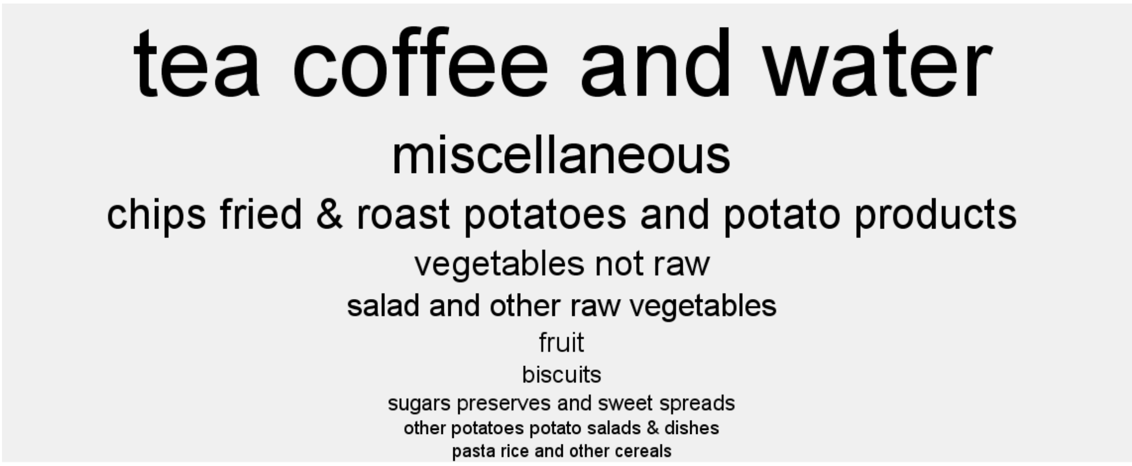


(g) Topic 7 : Lunch/sandwiches

(h) Topic 8 : Drinks and snacks

Figure C.10: NANS with NDNS RP food groups 10 Type model

(i) Topic 9 : Drinks and snacks

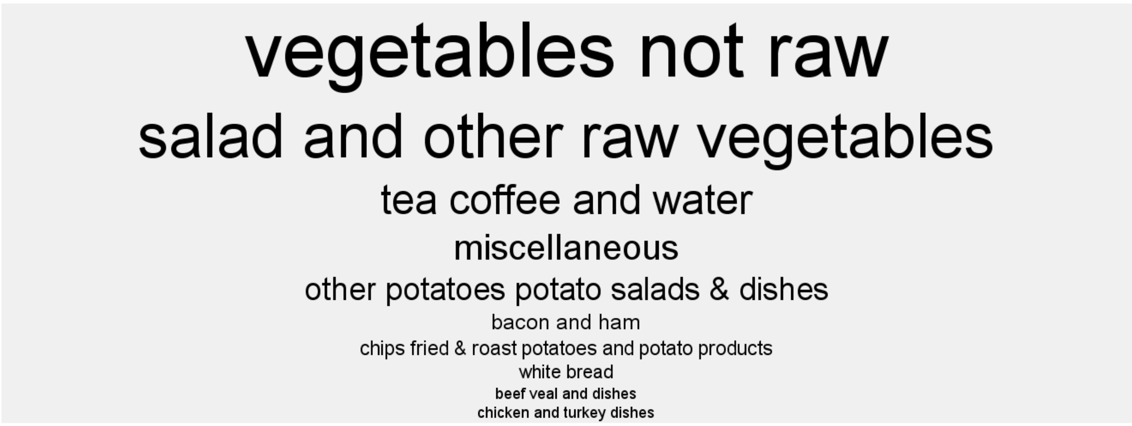


(j) Topic 10 : Main meal

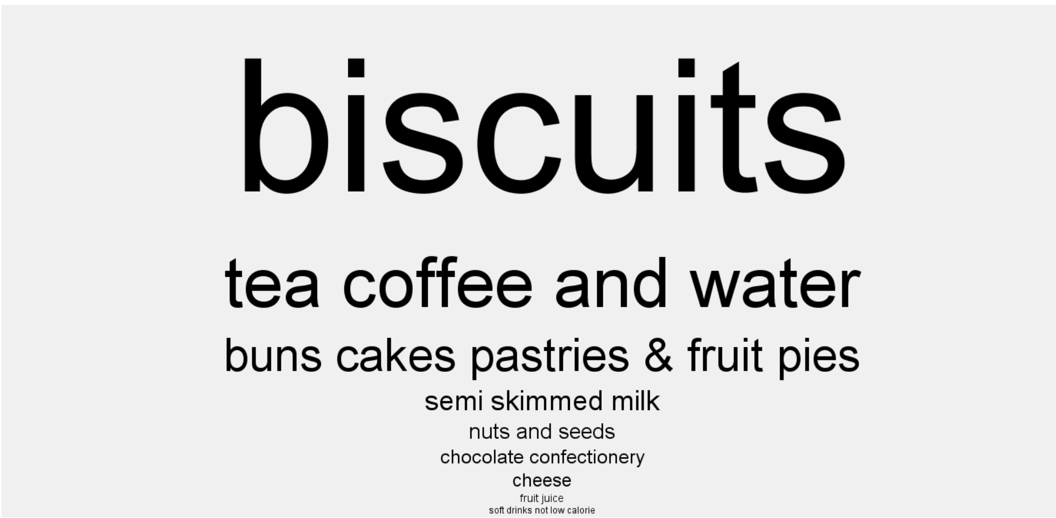
Figure C.10: NANS with NDNS RP food groups 10 Type model

C.11 NANS with NDNS food groups 15 Type Model

(a) Topic 1 : Drinks and snacks



(b) Topic 2 : Main meal



(c) Topic 3 : Drinks and snacks

Figure C.11: All 15 eating event types and labels NANS with NDNS RP food groups 15 Type model

(d) Topic 4 : Main meal

(e) Topic 5 : Breakfast

(f) Topic 6 : Drinks and snacks

Figure C.11: NANS with NDNS RP food groups 15 Type model

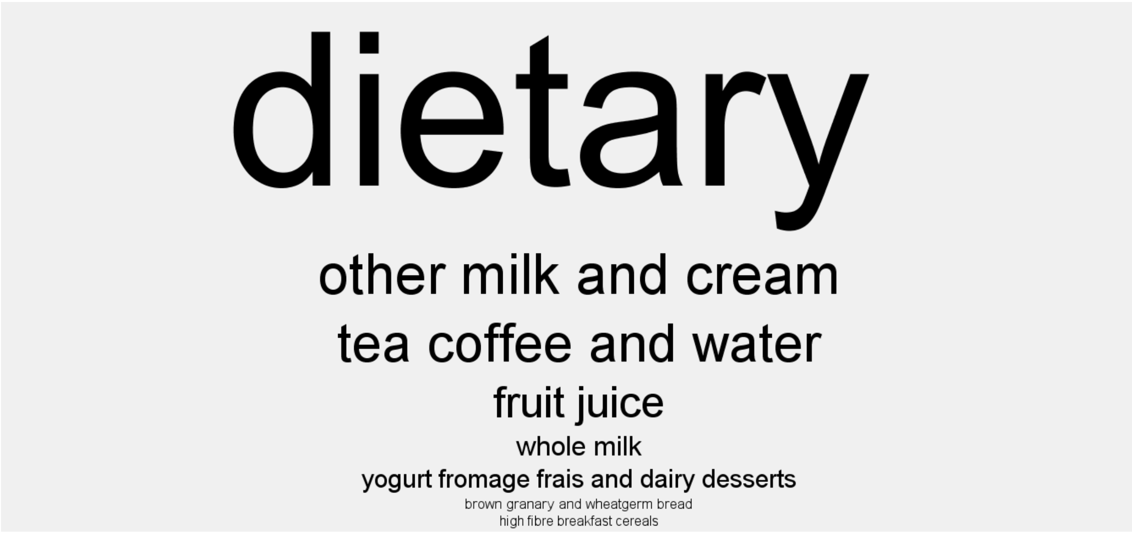
(g) Topic 7 : Breakfast

(h) Topic 8 : Light meal

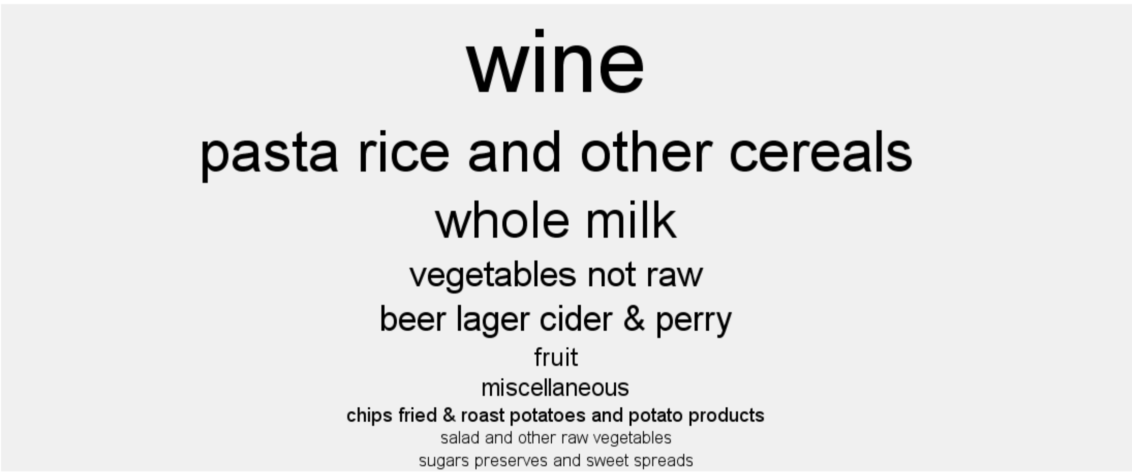
(i) Topic 9 : Drinks and snacks

Figure C.11: NANS with NDNS RP food groups 15 Type model

(j) Topic 10 : Drinks and snacks



(k) Topic 11 : Breakfast

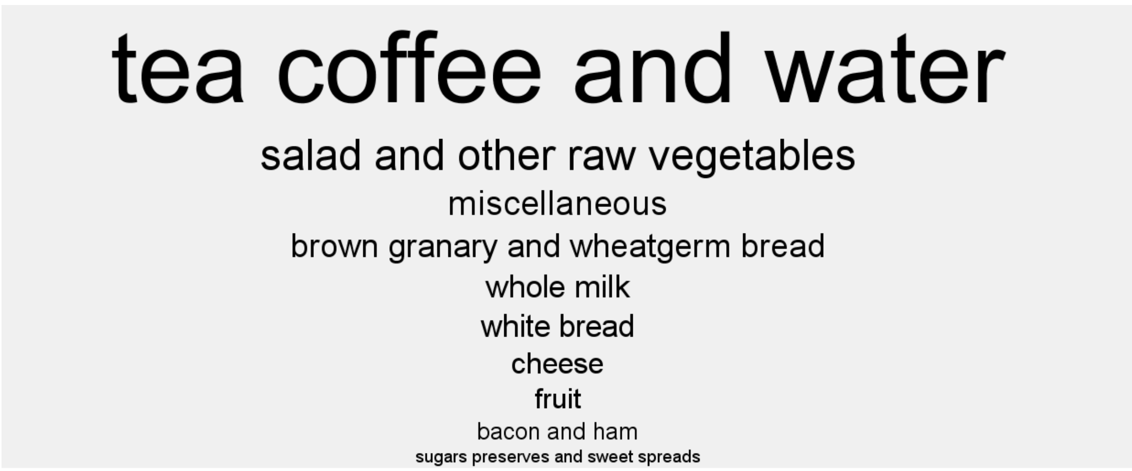


(l) Topic 12 : Light meal

Figure C.11: NANS with NDNS RP food groups 15 Type model

(m) Topic 13 : Main meal

(n) Topic 14 : Lunch/sandwiches



(o) Topic 15 : Lunch/sandwiches

Figure C.11: NANS with NDNS RP food groups 15 Type model

Bibliography

[1] W. H. Organization, “The World Health Report — Health systems financing: the path to universal coverage,” World Health Organization, Tech. Rep., 2010.

[2] A. Steventon, M. Bardsley, and J. Billings, “Effect of telehealth on use of secondary care and mortality : findings from the Whole System Demonstrator cluster randomised trial,” *BMJ*, vol. 3874, pp. 1–15, 2012.

[3] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers, “A review of wearable sensors and systems with application in rehabilitation.” *J. Neuroeng. Rehabil.*, vol. 9, no. 1, p. 21, Jan. 2012.

[4] N. Zhu, T. Diethe, M. Camplani, L. Tao, A. Burrows, N. Twomey, D. Kaleshi, M. Mirmehdi, P. Flach, and I. Craddock, “Bridging e-Health and the Internet of Things: The SPHERE Project,” *IEEE Intell. Syst.*, vol. 30, no. 4, pp. 39–46, Jul. 2015.

[5] Office for Budget Responsibility, “Fiscal Sustainability Report 2012,” Office for Budget Responsibility, Tech. Rep., 2012.

[6] ——, “Fiscal Sustainability Report 2015,” Office for Budget Responsibility, Tech. Rep., 2015.

[7] A. R. Hardisty, S. C. Peirce, A. Preece, C. E. Bolton, E. C. Conley, W. A. Gray, O. F. Rana, Z. Yousef, and G. Elwyn, “Bridging two translation gaps: a new informatics research agenda for telemonitoring of chronic disease.” *Int. J. Med. Inform.*, vol. 80, no. 10, pp. 734–44, Oct. 2011.

[8] S. S. Lim, *et al.*, “A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010,” *Lancet*, vol. 380, no. 9859, pp. 2224–2260, Dec. 2012.

[9] J. Spijker and J. MacInnes, “Population ageing: the timebomb that isn’t?” *BMJ*, vol. 347, no. 1, Jan. 2013.

[10] GBD 2015 Neurological Disorders Collaborator Group, V. L. Feigin, and T. Vos, "Global, regional, and national burden of neurological disorders during 1990-2015: a systematic analysis for the Global Burden of Disease Study 2015." *Lancet. Neurol.*, vol. 0, no. 0, Sep. 2017.

[11] J. L. Carús Candás, V. Peláez, G. López, M. Á. Fernández, E. Álvarez, and G. Díaz, "An automatic data mining method to detect abnormal human behaviour using physical activity measurements," *Pervasive Mob. Comput.*, vol. 15, pp. 228–241, Dec. 2014.

[12] G. Spina, P. Casale, P. S. Albert, J. Alison, J. Garcia-Aymerich, R. W. Costello, N. A. Hernandes, A. J. R. van Gestel, J. D. Leuppi, R. Mesquita, S. J. Singh, F. W. J. M. Smeenk, R. Tal-Singer, E. F. M. Wouters, M. A. Spruit, and A. C. den Brinker, "Identifying Physical Activity Profiles in COPD Patients Using Topic Models." *IEEE J. Biomed. Heal. informatics*, vol. 19, no. 5, pp. 1567–76, Sep. 2015.

[13] U. Ekelund, J. Luan, L. B. Sherar, D. W. Esliger, P. Griew, and A. Cooper, "Moderate to vigorous physical activity and sedentary time and cardiometabolic risk factors in children and adolescents." *JAMA*, vol. 307, no. 7, pp. 704–12, Feb. 2012.

[14] M. G. Davis, K. R. Fox, A. Stathi, T. Trayers, J. L. Thompson, and A. R. Cooper, "Objectively measured sedentary time and its association with physical function in older adults." *J. Aging Phys. Act.*, vol. 22, no. 4, pp. 474–81, Oct. 2014.

[15] D. J. Cook, N. C. Krishnan, and P. Rashidi, "Activity discovery and activity recognition: a new partnership." *IEEE Trans. Cybern.*, vol. 43, no. 3, pp. 820–8, Jun. 2013.

[16] N. Khanna, H. A. Eicher-Miller, C. J. Boushey, S. B. Gelfand, and E. J. Delp, "Temporal Dietary Patterns Using Kernel k-Means Clustering." in *IEEE Int. Symp. Multimed.*, Dec. 2011, pp. 375–380.

[17] D. Chapelot, "The role of snacking in energy balance: a biobehavioral approach." *J. Nutr.*, vol. 141, no. 1, pp. 158–62, Jan. 2011.

[18] R. Miller, B. Benelam, S. A. Stanner, and J. L. Buttriss, "Is snacking good or bad for health: An overview," *Nutr. Bull.*, vol. 38, no. 3, pp. 302–322, Sep. 2013.

[19] J. Synnott, C. Nugent, and P. Jeffers, "Simulation of Smart Home Activity Datasets." *Sensors (Basel).*, vol. 15, no. 6, pp. 14 162–79, Jan. 2015.

[20] G. Sprint, D. J. Cook, and D. L. Weeks, "Toward Automating Clinical Assessments: A Survey of the Timed Up and Go," *IEEE Rev. Biomed. Eng.*, vol. 8, pp. 64–77, 2015.

[21] P. Bonato, "Advances in wearable technology and applications in physical medicine and rehabilitation." *J. Neuroeng. Rehabil.*, vol. 2, no. 1, p. 2, Feb. 2005.

[22] I. Yamada and G. Lopez, "Wearable Sensing Systems for Healthcare Monitoring," in *Symp. VLSI Technol.*, Honolulu, HI, 2012, pp. 5 – 10.

[23] A. Pantelopoulos and N. Bourbakis, "A Survey on Wearable Sensor-Based Systems for Health Monitoring and Prognosis," *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.*, vol. 40, no. 1, pp. 1–12, Jan. 2010.

[24] R. Steele, A. Lo, C. Secombe, and Y. K. Wong, "Elderly persons' perception and acceptance of using wireless sensor networks to assist healthcare." *Int. J. Med. Inform.*, vol. 78, no. 12, pp. 788–801, Dec. 2009.

[25] A. P. Abidoye, "Using Wearable Sensors for Remote Healthcare Monitoring System," *J. Sens. Technol.*, vol. 01, no. 02, pp. 22–28, Jun. 2011.

[26] M. Alwan, S. Dalal, D. Mack, S. W. Kell, B. Turner, J. Leachtenauer, and R. Felder, "Impact of monitoring technology in assisted living: outcome pilot." *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 1, pp. 192–8, Jan. 2006.

[27] B. Logan, J. Healey, M. Philipose, E. M. Tapia, and S. Intille, "A long-term evaluation of sensing modalities for activity recognition," in *Proc. 9th Int. Conf. Ubiquitous Comput. UbiComp '07.* Springer-Verlag, Sep. 2007, pp. 483–500.

[28] A. A. Chaaraoui, J. R. Padilla-López, F. J. Ferrández-Pastor, M. Nieto-Hidalgo, and F. Flórez-Revuelta, "A vision-based system for intelligent monitoring: human behaviour analysis and privacy by context." *Sensors (Basel).*, vol. 14, no. 5, pp. 8895–925, Jan. 2014.

[29] University of Bristol, University of Reading, and University of Southampton, "SPHERE IRC Website," 2014. [Online]. Available: http://www.irc-sphere.ac.uk/

[30] I. Craddock, "SPHERE Proposal," University of Bristol, Tech. Rep., 2013.

[31] M. H. Rahmana, M. Pickering, D. Kerr, C. Boushey, and E. Delp, "A New Texture Feature for Improved Food Recognition Accuracy in a Mobile Phone Based Dietary Assessment System," in *2012 IEEE Int. Conf. Multimed. Expo Work.* IEEE, Jul. 2012, pp. 418–423.

[32] T. Huynh, M. Fritz, and B. Schiele, "Discovery of activity patterns using topic models," in *Proc. 10th Int. Conf. Ubiquitous Comput.*, New York, New York, USA, Sep. 2008, p. 10.

[33] R. White, W. S. Harwin, W. Holderbaum, and L. Johnson, "Investigating Eating Behaviours Using Topic Models," in *2015 IEEE 14th Int. Conf. Mach. Learn. Appl.* IEEE, Dec. 2015, pp. 265–270.

[34] R. C. King, E. Villeneuve, R. J. White, R. S. Sherratt, W. Holderbaum, and W. S. Harwin, "Application of data fusion techniques and technologies for wearable health monitoring," *Med. Eng. Phys.*, vol. 42, pp. 1–12, Apr. 2017.

[35] E. Villeneuve, W. Harwin, W. Holderbaum, R. S. Sherratt, and R. White, "Signal Quality and Compactness of a Dual-Accelerometer System for Gyro-Free Human Motion Analysis," *IEEE Sens. J.*, vol. 16, no. 16, pp. 6261–6269, Aug. 2016.

[36] D. Blei, L. Carin, and D. Dunson, "Probabilistic Topic Models: A focus on graphical model design and applications to document and image analysis." *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 55–65, Nov. 2010.

[37] D. Koller and N. Friedman, *Probabilistic Graphical Models: principles and techniques.* MIT Press, 2009.

[38] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer International Publishing, 2006.

[39] B. A. Frigyik, A. Kapila, and M. R. Gupta, "Introduction to the Dirichlet Distribution and Related Processes - UWEE Tech Report Series," Department of Electrical Engineering, University of Washington, Seattle, WA, USA, Tech. Rep., 2010.

[40] S. Tu, "The Dirichlet-Multinomial and Dirichlet-Categorical models for Bayesian inference," Computer Science Division, UC Berkeley, Tech. Rep., 2014.

[41] D. M. Blei, "Topic Models (Lecture Notes)," Cambridge, UK, 2009.

[42] The European Mathematical Society, "Atom," 2016. [Online]. Available: https://www.encyclopediaofmath.org/index.php/Atom

[43] M. J. Wainwright and M. I. Jordan, "Graphical Models, Exponential Families, and Variational Inference," *Found. Trends Mach. Learn.*, vol. 1, no. 1–2, pp. 1–305, Nov. 2008.

[44] Z. Ghahramani, “Introduction to Graphical Models (Lecture Notes),” Cambridge, UK, 2009. [Online]. Available: http://videolectures.net/mlss09uk{_}ghahramani{_}gm/

[45] M. I. Jordan, “Graphical Models,” *Stat. Sci.*, vol. 19, no. 1, pp. 140–155, Feb. 2004.

[46] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[47] D. M. Blei and J. D. Lafferty, “Topic models,” in *Text Min. Classif. Clust. Appl.*, A. Srivastava and M. Sahami, Eds. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series, 2009.

[48] D. M. Blei, “Probabilistic topic models,” *Commun. ACM*, vol. 55, no. 4, p. 77, Apr. 2012.

[49] ——, “Probabilistic Topic Models (MLSS 2012 Lecture Notes),” 2012.

[50] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, “On Smoothing and Inference for Topic Models,” in *Proc. Twenty-Fifth Conf. Uncertain. Artif. Intell.*, ser. UAI ’09. Arlington, Virginia, United States: AUAI Press, 2009, pp. 27–34.

[51] D. M. Blei, “lda.c program,” 2006. [Online]. Available: http://www.cs.princeton.edu/{~}blei/lda-c/

[52] ——, “Variational Inference (Lecture Notes),” 2011. [Online]. Available: https://www.cs.princeton.edu/courses/archive/fall11/cos597C/lectures/variational-inference-i.pdf

[53] D. M. Blei and J. D. Lafferty, “Dynamic topic models,” in *Proc. 23rd Int. Conf. Mach. Learn. - ICML ’06*. New York, New York, USA: ACM Press, Jun. 2006, pp. 113–120.

[54] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, “The author-topic model for authors and documents,” in *UAI ’04 Proc. 20th Conf. Uncertain. Artif. Intell.* AUAI Press, Jul. 2004, pp. 487–494.

[55] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, “Evaluation methods for topic models,” in *Proc. 26th Annu. Int. Conf. Mach. Learn. - ICML ’09*. New York, New York, USA: ACM Press, Jun. 2009, pp. 1–8.

[56] D. Newman, S. Karimi, and L. Cavedon, “External evaluation of topic models,” in *Proc. 14th Australas. Doc. Comput. Symp.* Sydney, Australia: School of Information Technologies, University of Sydney, 2009, pp. 1–8.

[57] J. Chang, J. L. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, "Reading Tea Leaves: How Humans Interpret Topic Models." in *Adv. Neural Inf. Process. Syst. 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Curran Associates, Inc., Jan. 2009, pp. 288–296.

[58] H. Chan and L. Akoglu, "External Evaluation of Topic Models: A Graph Mining Approach," in *2013 IEEE 13th Int. Conf. Data Min.* IEEE, Dec. 2013, pp. 973–978.

[59] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Hum. Lang. Technol. 2010 Annu. Conf. North Am. Chapter Assoc. Comput. Linguist.* Association for Computational Linguistics, Jun. 2010, pp. 100–108.

[60] C. Chen, R. Jafari, and N. Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition," *Multimed. Tools Appl.*, pp. 1–21, Dec. 2015.

[61] S. A. Lowe and G. ÓLaighin, "Monitoring human health behaviour in one's living environment: A technological review," *Med. Eng. Phys.*, vol. 36, no. 2, pp. 147–168, 2014.

[62] Q. Ni, A. B. G. Hernando, and I. P. de la Cruz, "The elderly's independent living in smart homes: A characterization of activities and sensing infrastructure survey to facilitate services development," *Sensors (Switzerland)*, vol. 15, no. 5, pp. 11 312–11 362, May 2015.

[63] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-based activity recognition," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 6, pp. 790–808, Nov. 2012.

[64] A. Mirza, "Data Fusion Architectures for Sensor Platforms," in *Aerosp. Conf. 2008 IEEE*, Big Sky, MT, 2008, pp. 1 – 13.

[65] L. Wald, "Some terms of reference in data fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1190–1193, May 1999.

[66] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Inf. Fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[67] F. Castanedo, "A review of data fusion techniques." *Sci. World J.*, vol. 2013, no. 704504, p. 19, 2013.

[68] P. Varshney, "Multisensor data fusion," *Electron. Commun. Eng. J.*, vol. 9, no. 6, pp. 245–253, Dec. 1997.

[69] J. Esteban, A. Starr, R. Willetts, P. Hannah, and P. Bryanston-Cross, "A review of data fusion models and architectures: towards engineering guidelines," *Neural Comput. Appl.*, vol. 14, no. 4, pp. 273–281, Jun. 2005.

[70] D. Hall and J. Llinas, "An introduction to multisensor data fusion," *Proc. IEEE*, vol. 85, no. 1, pp. 6–23, 1997.

[71] M. Kokar and K. Kim, "Review of multisensor data fusion architectures and techniques," in *Proc. 8th IEEE Int. Symp. Intell. Control.* IEEE, 1993, pp. 261–266.

[72] W. Koch, F. Fkie, and A. J. D. L. Level, "The JDL Model of Data Fusion Applied to Cyber-Defence – a Review Paper," in *2012 Work. Sens. Data Fusion Trends, Solut. Appl.*, no. September, Bonn, 2012, pp. 4–6.

[73] D. Hall and J. Llinas, Eds., *Handbook of Multisensor Data Fusion.* CRC Press, 2001.

[74] R. Luo and M. Kay, "Multisensor integration and fusion in intelligent systems," *IEEE Trans. Syst. Man. Cybern.*, vol. 19, no. 5, pp. 901–931, 1989.

[75] R. C. Luo, C. C. Chang, and C. C. Lai, "Multisensor Fusion and Integration : Theories , Applications , and its Perspectives," *IEEE Sens. J.*, vol. 11, no. 12, pp. 3122–3138, 2011.

[76] B. Dasarathy, "Sensor fusion potential exploitation-innovative architectures and illustrative applications," *Proc. IEEE*, vol. 85, no. 1, pp. 24–38, 1997.

[77] H. Lee, K. Park, B. Lee, J. Choi, and R. Elmasri, "Issues in data fusion for healthcare monitoring," in *Proc. 1st ACM Int. Conf. PErvasive Technol. Relat. to Assist. Environ. - PETRA '08.* New York, New York, USA: ACM Press, 2008, p. 1.

[78] J. Gong, L. Cui, K. Xiao, and R. Wang, "MPD-Model: A Distributed Multipreference-Driven Data Fusion Model and Its Application in a WSNs-Based Healthcare Monitoring System," *Int. J. Distrib. Sens. Networks*, vol. 2012, pp. 1–13, 2012.

[79] G. Fortino, S. Galzarano, R. Gravina, and W. Li, "A framework for collaborative computing and multi-sensor data fusion in body sensor networks," *Inf. Fusion*, vol. 22, pp. 50–70, 2015.

[80] S. Rodríguez, J. F. De Paz, G. Villarrubia, C. Zato, J. Bajo, and J. M. Corchado, "Multi-Agent Information Fusion System to manage data from a WSN in a residential home," *Inf. Fusion*, vol. 23, pp. 43–57, 2015.

[81] H. Cao, M. N. Nguyen, C. Phua, S. Krishnaswamy, and X.-L. Li, "An integrated framework for human activity classification," in *Proc. 2012 ACM Conf. Ubiquitous Comput. - UbiComp '12.* New York, New York, USA: ACM Press, 2012, p. 331.

[82] X. Lai, Q. Liu, X. Wei, W. Wang, G. Zhou, and G. Han, "A Survey of Body Sensor Networks," *Sensors*, vol. 13, no. 5, pp. 5406–5447, Apr. 2013.

[83] I. Pires, N. Garcia, N. Pombo, and F. Flórez-Revuelta, "From Data Acquisition to Data Fusion: A Comprehensive Review and a Roadmap for the Identification of Activities of Daily Living Using Mobile Devices," *Sensors*, vol. 16, no. 2, p. 184, Feb. 2016.

[84] G. Koshmak, A. Loutfi, M. Linden, G. Koshmak, A. Loutfi, and M. Linden, "Challenges and Issues in Multisensor Fusion Approach for Fall Detection: Review Paper," *J. Sensors*, vol. 2016, pp. 1–12, 2016.

[85] M. R. Alam, M. B. I. Reaz, and M. A. M. Ali, "A review of smart homes - Past, present, and future," *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.*, vol. 42, no. 6, pp. 1190–1203, Nov. 2012.

[86] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *IEEE J. Biomed. Heal. Informatics*, vol. 17, no. 3, pp. 579–590, May 2013.

[87] D. Ding, R. A. Cooper, P. F. Pasquina, and L. Fici-Pasquina, "Sensor technology for smart homes." *Maturitas*, vol. 69, no. 2, pp. 131–6, Jun. 2011.

[88] J. Favela, J. Kaye, M. Skubic, M. Rantz, and M. Tentori, "Living Labs for Pervasive Healthcare Research," *IEEE Pervasive Comput.*, vol. 14, no. 2, pp. 86–89, Apr. 2015.

[89] M. Ogawa, R. Suzuki, S. Otake, T. Izutsu, T. Iwaya, and T. Togawa, "Long term remote behavioral monitoring of elderly by using sensors installed in ordinary houses," in *Proc. 2nd Annu. Int. IEEE-EMBS Spec. Top. Conf. Microtechnologies Med. Biol.*, 2002, pp. 322–325.

[90] T. Zhao, H. Ni, X. Zhou, L. Qiang, D. Zhang, and Z. Yu, "Detecting Abnormal Patterns of Daily Activities for the Elderly Living Alone," in *Heal. Inf. Sci. Third Int. Conf. HIS 2014*, ser. Lecture Notes in Computer Science, Y. Zhang,

G. Yao, J. He, L. Wang, N. R. Smalheiser, and X. Yin, Eds., vol. 8423. Shenzhen, China: Springer International Publishing, 2014, pp. 95 – 108.

[91] T. L. M. van Kasteren, G. Englebienne, and B. J. A. Kröse, "Activity recognition using semi-Markov models on real world smart home datasets," *J. Ambient Intell. Smart Environ.*, vol. 2, no. 3, pp. 311–325, Aug. 2010.

[92] L. G. Fahad, A. Ali, and M. Rajarajan, "Long term analysis of daily activities in a smart home," in *Eur. Symp. Artif. Neural Networks, Comput. Intell. Mach. Learn.*, 2013, pp. 419 – 424.

[93] M. Alwan, S. Member, D. C. Mack, S. Dalal, S. Kell, B. Turner, and R. A. Felder, "Impact of Passive In-Home Health Status Monitoring Technology in Home Health: Outcome Pilot 1," in *Proc. 1st Distrib. Diagnosis Home Healthc. Conf.*, Arlington, Virginia, 2006, pp. 2–5.

[94] D. J. Cook, A. S. Crandall, B. L. Thomas, and N. C. Krishnan, "CASAS: A smart home in a box," *Computer (Long. Beach. Calif).*, vol. 46, no. 7, pp. 62–69, Jul. 2013.

[95] S. S. Intille, K. Larson, J. Beaudin, J. Nawyn, E. Munguia Tapia, and P. Kaushik, "A living laboratory for the design and evaluation of ubiquitous computing technologies," in *Ext. Abstr. 2005 Conf. Hum. Factors Comput. Syst.*, 2005, pp. 1941–1944.

[96] P. Rashidi, D. J. Cook, L. B. Holder, and M. Schmitter-Edgecombe, "Discovering Activities to Recognize and Track in a Smart Environment." *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 4, pp. 527–539, Jan. 2011.

[97] T. van Kasteren, A. Noulas, G. Englebienne, and B. Kröse, "Accurate activity recognition in a home setting," in *Proc. 10th Int. Conf. Ubiquitous Comput. - UbiComp '08.* New York, New York, USA: ACM Press, Sep. 2008, p. 1.

[98] T. van Kasteren, G. Englebienne, and B. Krose, "Human activity recognition from wireless sensor network data - benchmark and software," in *Act. Recognit. Pervasive Intell. Environ.*, L. Chen, C. Nugent, J. Biswas, and J. Hoey, Eds. Atlantis Press, 2010.

[99] F. J. Ordonez, G. Englebienne, P. de Toledo, T. van Kasteren, A. Sanchis, and B. Krose, "In-Home Activity Recognition: Bayesian Inference for Hidden Markov Models," *IEEE Pervasive Comput.*, vol. 13, no. 3, pp. 67–75, Jul. 2014.

[100] R. F. Dickerson, E. I. Gorlin, and J. A. Stankovic, "Empath: a continuous remote emotional health monitoring system for depressive illness," in *Proc. 2nd*

Conf. Wirel. Heal. - WH '11. New York, New York, USA: ACM Press, 2011, p. 1.

[101] O. Lara and M. Labrador, "A Survey on Human Activity Recognition using Wearable Sensors," *IEEE Commun. Surv. Tutorials*, vol. 15, no. 3, pp. 1192–1209, Jan. 2013.

[102] R. Khusainov, D. Azzi, I. E. Achumba, and S. D. Bersch, "Real-time human ambulation, activity, and physiological monitoring: taxonomy of issues, techniques, applications, challenges and limitations." *Sensors (Basel).*, vol. 13, no. 10, pp. 12 852–12 902, Sep. 2013.

[103] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Comput. Surv.*, vol. 46, no. 3, pp. 1–33, Feb. 2014.

[104] A. Godfrey, R. Conway, D. Meagher, and G. ÓLaighin, "Direct measurement of human movement by accelerometry," *Med. Eng. Phys.*, vol. 30, no. 10, pp. 1364–1386, 2008.

[105] A. Avci, S. Bosch, M. Marin-perianu, R. Marin-perianu, and P. Havinga, "Activity recognition using inertial sensing for healthcare, wellbeing and sports applications: A survey," in *ARCS Work.*, 2010, pp. 167–176.

[106] S. J. Preece, J. Y. Goulermas, L. P. J. Kenney, and D. Howard, "A comparison of feature extraction methods for the classification of dynamic activities from accelerometer data." *IEEE Trans. Biomed. Eng.*, vol. 56, no. 3, pp. 871–9, Mar. 2009.

[107] S. J. Preece, J. Y. Goulermas, L. P. J. Kenney, D. Howard, K. Meijer, and R. Crompton, "Activity identification using body-mounted sensors–a review of classification techniques." *Physiol. Meas.*, vol. 30, no. 4, pp. R1–33, Apr. 2009.

[108] L. C. Jatoba, U. Grossmann, C. Kunze, J. Ottenbacher, and W. Stork, "Context-aware mobile health monitoring: Evaluation of different pattern recognition methods for classification of physical activity," in *2008 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* IEEE, Aug. 2008, pp. 5250–5253.

[109] A. Mannini and A. M. Sabatini, "Machine learning methods for classifying human physical activity from on-body accelerometers," *Sensors*, vol. 10, no. 2, pp. 1154–1175, Feb. 2010.

[110] L. Bao and S. Intille, "Activity Recognition from User-Annotated Acceleration Data," in *Pervasive Comput. SE - 1*, ser. Lecture Notes in Computer Science,

A. Ferscha and F. Mattern, Eds. Springer Berlin Heidelberg, 2004, vol. 3001, pp. 1–17.

[111] J. Wang, R. Chen, X. Sun, M. F. She, and Y. Wu, "Recognizing Human Daily Activities From Accelerometer Signal," *Procedia Eng.*, vol. 15, pp. 1780–1786, 2011.

[112] K. K. B. Peetoom, M. A. S. Lexis, M. Joore, C. D. Dirksen, and L. P. De Witte, "Literature review on monitoring technology and their outcomes in independently living elderly people," *Disabil. Rehabil. Assist. Technol.*, vol. 10, no. 4, pp. 271 – 294, Jul. 2015.

[113] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Forster, G. Troster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, and J. d. R. Millan, "Collecting complex activity datasets in highly rich networked sensor environments," in *2010 Seventh Int. Conf. Networked Sens. Syst.* IEEE, Jun. 2010, pp. 233–240.

[114] R. Chavarriaga, H. Sagha, A. Calatroni, S. T. Digumarti, G. Tröster, J. D. R. Millán, and D. Roggen, "The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition," *Pattern Recognit. Lett.*, vol. 34, no. 15, pp. 2033–2042, 2013.

[115] N. Twomey, T. Diethe, M. Kull, H. Song, M. Camplani, S. Hannuna, X. Fafoutis, N. Zhu, P. Woznowski, P. Flach, and I. Craddock, "The SPHERE Challenge: Activity Recognition with Multimodal Sensor Data," *CoRR*, vol. abs/1603.0, 2016.

[116] X. Liu, L. Liu, and P.-N. Tan, "Location based Hierarchical Approach for Activity Recognition with Multi-modal Sensor Data," in *26th Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discov.*, Riva del Garda, 2016.

[117] G. Virone, M. Alwan, S. Dalal, S. W. Kell, B. Turner, J. A. Stankovic, and R. Felder, "Behavioral patterns of older adults in assisted living," *IEEE Trans. Inf. Technol. Biomed.*, vol. 12, no. 3, pp. 387–398, May 2008.

[118] S. Ranasinghe, F. A. Machot, and H. C. Mayr, "A Review on Applications of Activity Recognition Systems with Regard to Performance and Evaluation," *Int. J. Distrib. Sens. Networks*, vol. 12, no. 8, pp. 1–21, Aug. 2016.

[119] L. Ferrari and M. Mamei, "Discovering daily routines from Google Latitude with topic models," in *2011 IEEE Int. Conf. Pervasive Comput. Commun. Work.*, Mar. 2011, pp. 432–437.

[120] K. Farrahi and D. Gatica-Perez, “Discovering routines from large-scale human locations using probabilistic topic models,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 1, pp. 1–27, Jan. 2011.

[121] ——, “Discovering human routines from cell phone data with topic models,” *Proc. - Int. Symp. Wearable Comput. ISWC*, pp. 29–32, 2008.

[122] B. Nabaei and M. Ester, “Activity Monitoring Using Topic Models,” in *IEEE Conf. Intell. Secur. Informatics*, 2016, pp. 115–120.

[123] U. Steinhoff and B. Schiele, “An Exploration of Daily Routine Modeling Based on Bluetooth and GSM-Data,” in *Int. Symp. Wearable Comput.*, Sep. 2009, pp. 141–142.

[124] J. Seiter, O. Amft, and G. Tröster, “Assessing Topic Models: How to Obtain Robustness?” *AwareCast 2012 Work. Recent Adv. Behav. Predict. Pro-active Pervasive Comput.*, 2012.

[125] J. Seiter, O. Amft, M. Rossi, and G. Tröster, “Discovery of activity composites using topic models: An analysis of unsupervised methods,” *Pervasive Mob. Comput.*, vol. 15, pp. 215–227, 2014.

[126] J. Seiter, A. Derungs, C. Schuster-Amft, O. Amft, and G. Tröster, “Daily life activity routine discovery in hemiparetic rehabilitation patients using topic models,” *Methods Inf. Med.*, vol. 54, no. 3, pp. 248–255, Feb. 2015.

[127] S. Kim, M. Li, S. Lee, U. Mitra, A. Emken, D. Spruijt-Metz, M. Annavaram, and S. Narayanan, “Modeling high-level descriptions of real-life physical activities using latent topic modeling of multimodal sensor signals,” in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, 2011, pp. 6033–6036.

[128] L. Peng, L. Chen, X. Wu, H. Guo, and G. Chen, “Hierarchical Complex Activity Representation and Recognition using Topic Model and Classifier Level Fusion,” *IEEE Trans. Biomed. Eng.*, vol. PP, no. 99, 2016.

[129] F.-T. Sun, H.-T. Cheng, C. Kuo, and M. Griss, “Nonparametric discovery of human routines from sensor data,” in *IEEE Int. Conf. Pervasive Comput. Commun.*, Mar. 2014, pp. 11–19.

[130] T. Nguyen, S. K. Gupta, D. Phung, and S. Venkatesh, “A Bayesian Nonparametric Framework for Activity Recognition using Accelerometer Data,” in *Int. Conf. Pattern Recognit.*, Aug. 2014.

[131] Y. Zhu, Y. Arase, X. Xie, and Q. Yang, "Bayesian nonparametric modeling of user activities," in *Proc. 2011 Int. Work. Trajectory data Min. Anal.* New York, New York, USA: ACM Press, 2011, pp. 1–4.

[132] E. Rogers, J. D. Kelleher, and R. J. Ross, "Using topic modelling algorithms for hierarchical activity discovery," *Adv. Intell. Syst. Comput.*, vol. 476, pp. 41–48, 2016.

[133] K. Rieping, G. Englebienne, and B. Kröse, "Behavior analysis of elderly using topic models," *Pervasive Mob. Comput.*, vol. 15, pp. 181–199, 2014.

[134] F. Castanedo, D. L. De-Ipiña, H. K. Aghajan, and R. Kleihorst, "Learning routines over long-term sensor data using topic models," *Expert Syst.*, vol. 31, no. 4, pp. 365–377, Sep. 2014.

[135] B. Chikhaoui, S. Wang, and H. Pigot, "ADR-SPLDA: Activity discovery and recognition by combining sequential patterns and latent Dirichlet allocation," *Pervasive Mob. Comput.*, vol. 8, no. 6, pp. 845–862, 2012.

[136] Y. Chen, T. Diethe, and P. Flach, "ADL^{TM}: A Topic Model for Recognition of Activities of Daily Living in a SmartHome," in *Proc. Twenty-Fifth Int. Jt. Conf. Artif. Intell.*, S. Kambhampati, Ed. AAAI Press, 2016, pp. 1404–1410.

[137] A. Rai, Z. Yan, D. Chakraborty, T. K. Wijaya, and K. Aberer, "Mining complex activities in the wild via a single smartphone accelerometer," in *Proc. Sixth Int. Work. Knowl. Discov. from Sens. Data - SensorKDD '12.* New York, New York, USA: ACM Press, 2012, pp. 43–51.

[138] I. Kennedy, U. Naeem, and A.-R. Tawil, "A dynamic segmentation based activity discovery through topic modelling," in *IET Int. Conf. Technol. Act. Assist. Living.* London: IET Digital Library, Jan. 2015, pp. 1–6.

[139] I. K. Ihianle, U. Naeem, and A.-R. Tawil, "Recognition of Activities of Daily Living from Topic Model," *Procedia Comput. Sci.*, vol. 98, pp. 24–31, Dec. 2016.

[140] J. Zheng, S. Liu, and L. M. Ni, "Effective Routine Behavior Pattern Discovery from Sparse Mobile Phone Data via Collaborative Filtering," in *2013 IEEE Int. Conf. Pervasive Comput. Commun.*, 2013, pp. 29–37.

[141] N. Eagle and A. Pentland, "Reality mining: Sensing complex social systems," *Pers. Ubiquitous Comput.*, vol. 10, no. 4, pp. 255–268, May 2006.

[142] N. Eagle and A. S. Pentland, "Eigenbehaviors: identifying structure in routine," *Behav. Ecol. Sociobiol.*, vol. 63, no. 7, pp. 1057–1066, Apr. 2009.

[143] M. A. Azam and L. Tokarchuk, "Behaviour detection using bluetooth proximity data," in *Proc. Netw. Electron. Commer. Res. Conf.*, 2009.

[144] M. A. Azam, J. Loo, A. Lasebae, S. K. A. Khan, and W. Ejaz, "Behavioural analysis of low entropy mobile people using contextual information," in *2012 IEEE Int. Conf. Pervasive Comput. Commun. Work.* IEEE, Mar. 2012, pp. 590–595.

[145] T. Barger, D. Brown, and M. Alwan, "Health-status monitoring through analysis of behavioral patterns," *IEEE Trans. Syst. Man*, vol. 35, pp. 22–27, 2005.

[146] C. Li and W. K. Cheung, "Recovering human mobility flow models and daily routine patterns in a smart environment," in *IEEE Int. Conf. Data Min. Work. ICDMW*, 2014, pp. 541–548.

[147] J. Yin, Q. Zhang, and M. Karunanithi, "Unsupervised daily routine and activity discovery in smart homes," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS*, 2015, pp. 5497–5500.

[148] Q. Lin, D. Zhang, D. Li, H. Ni, and X. Zhou, "Extracting Intra- and Inter-activity Association Patterns from Daily Routines of Elders," in *Incl. Soc. Heal. Wellbeing Community, Care Home 11th Int. Conf. Smart Homes Heal. Telemat. ICOST 2013, Singapore, June 19-21, 2013. Proc.*, J. Biswas, H. Kobayashi, L. Wong, B. Abdulrazak, and M. Mokhtari, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 36–44.

[149] P. Rashidi and D. J. Cook, "Mining sensor streams for discovering human activity patterns over time," in *Proc. - IEEE Int. Conf. Data Mining, ICDM.* IEEE, Dec. 2010, pp. 431–440.

[150] U. Blanke and B. Schiele, "Daily routine recognition through activity spotting," in *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics).* Springer Berlin Heidelberg, 2009, vol. 5561 LNCS, pp. 192–206.

[151] A. Derungs, J. Seiter, C. Schuster-Amft, and O. Amft, "Activity Patterns in Stroke Patients - Is There a Trend in Behaviour During Rehabilitation?" in *Hum. Behav. Underst.*, ser. Lecture Notes in Computer Science, A. A. Salah, B. J. Kröse, and D. J. Cook, Eds. Springer International Publishing, Jan. 2015, vol. 9277, pp. 146–159.

[152] S. Robben, M. Pol, and B. Kröse, "Longitudinal ambient sensor monitoring for functional health assessments," in *Proc. 2014 ACM Int. Jt. Conf. Pervasive*

Ubiquitous Comput. Adjun. Publ. New York, New York, USA: ACM Press, 2014, pp. 1209–1216.

[153] S. Robben, G. Englebienne, and B. Krose, "Delta Features From Ambient Sensor Data are Good Predictors of Change in Functional Health," *IEEE J. Biomed. Heal. Informatics*, vol. 21, no. 4, pp. 986–993, 2016.

[154] A. Wilbik, J. M. Keller, and G. L. Alexander, "Linguistic summarization of sensor data for eldercare," in *IEEE Int. Conf. Syst. Man, Cybern.*, Anchorage, AK, 2011, pp. 2595–2599.

[155] S. Wang, M. Skubic, and Y. Zhu, "Activity Density Map Visualization and Dissimilarity Comparison for Eldercare Monitoring," *IEEE Trans. Inf. Technol. Biomed.*, vol. 16, no. 4, pp. 607–614, Jul. 2012.

[156] M. Skubic, R. D. Guevara, and M. Rantz, "Automated Health Alerts Using In-Home Sensor Data for Embedded Health Assessment," *IEEE J. Transl. Eng. Heal. Med.*, vol. 3, pp. 1–11, 2015.

[157] P. N. Dawadi, D. J. Cook, and M. Schmitter-Edgecombe, "Modeling Patterns of Activities using Activity Curves." *Pervasive Mob. Comput.*, vol. 28, pp. 51–68, Jun. 2016.

[158] D. Elbert, H. Storf, M. Eisenbarth, O. Unalan, and M. Schmitt, "An approach for detecting deviations in daily routine for long-term behavior analysis," in *5th Int. Conf. Pervasive Comput. Technol. Healthc.*, Dublin, 2011, pp. 426–433.

[159] C. J. Boushey, D. A. Kerr, J. Wright, K. D. Lutes, D. S. Ebert, and E. J. Delp, "Use of technology in children's dietary assessment," *Eur. J. Clin. Nutr.*, vol. 63, pp. S50–S57, Feb. 2009.

[160] P. J. Stumbo, "New technology in dietary assessment: a review of digital methods in improving food record accuracy," *Proc. Nutr. Soc.*, vol. 72, no. 01, pp. 70–76, Feb. 2013.

[161] I. Elmadfa and A. L. Meyer, "Importance of food composition data to nutrition and public health," *Eur. J. Clin. Nutr.*, vol. 64, pp. S4–S7, Nov. 2010.

[162] T. A. Nicklas, T. Baranowski, K. W. Cullen, and G. Berenson, "Eating Patterns, Dietary Quality and Obesity," *J. Am. Coll. Nutr.*, vol. 20, no. 6, pp. 599–608, Dec. 2001.

[163] P. F. Jacques and K. L. Tucker, "Are dietary patterns useful for understanding the role of diet in chronic disease?" *Am. J. Clin. Nutr.*, vol. 73, no. 1, pp. 1–2, Jan. 2001.

[164] E. Erber, B. N. Hopping, A. Grandinetti, S.-Y. Park, L. N. Kolonel, and G. Maskarinec, "Dietary patterns and risk for diabetes: the multiethnic cohort." *Diabetes Care*, vol. 33, no. 3, pp. 532–8, Mar. 2010.

[165] R. F. Tayyem, H. A. Bawadi, I. Shehadah, L. M. Agraib, S. S. AbuMweis, T. Al-Jaberi, M. Al-Nusairr, K. E. Bani-Hani, and D. D. Heath, "Dietary patterns and colorectal cancer," *Clin. Nutr.*, vol. 36, no. 3, pp. 848–852, 2017.

[166] Y. He, C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp, "Analysis of food images: Features and classification," in *2014 IEEE Int. Conf. Image Process.* IEEE, Oct. 2014, pp. 2744–2748.

[167] G. Ciocca, P. Napoletano, and R. Schettini, "Food Recognition: A New Dataset, Experiments, and Results," *IEEE J. Biomed. Heal. Informatics*, vol. 21, no. 3, pp. 588–598, May 2017.

[168] J. Dehais, M. Anthimopoulos, S. Shevchik, and S. Mougiakakou, "Two-View 3D Reconstruction for Food Volume Estimation," *IEEE Trans. Multimed.*, vol. 19, no. 5, pp. 1090–1099, May 2017.

[169] C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, Y. Ma, S. Chen, and P. Hou, "A New Deep Learning-based Food Recognition System for Dietary Assessment on An Edge Computing Service Infrastructure," *IEEE Trans. Serv. Comput.*, pp. 1–1, 2017.

[170] O. Amft, "Ambient, On-Body, and Implantable Monitoring Technologies to Assess Dietary Behavior," in *Handb. Behav. Food Nutr.* New York, NY: Springer New York, 2011, pp. 3507–3526.

[171] B. Zhou, J. Cheng, P. Lukowicz, A. Reiss, and O. Amft, "Monitoring Dietary Behavior with a Smart Dining Tray," *IEEE Pervasive Comput.*, vol. 14, no. 4, pp. 46–56, Oct. 2015.

[172] S. Manton, G. Magerowski, L. Patriarca, and M. Alonso-Alonso, "The 'Smart Dining Table': Automatic Behavioral Tracking of a Meal with a Multi-Touch-Computer," *Front. Psychol.*, vol. 7, p. 142, Feb. 2016.

[173] B. Zhou, J. Cheng, M. Sundholm, A. Reiss, W. Huang, O. Amft, and P. Lukowicz, "Smart table surface: A novel approach to pervasive dining monitoring," in *2015 IEEE Int. Conf. Pervasive Comput. Commun.*, Mar. 2015, pp. 155–162.

[174] R. S. Mattfeld, E. R. Muth, and A. Hoover, "Measuring the consumption of individual solid and liquid bites using a table embedded scale during unrestricted eating," *IEEE J. Biomed. Heal. Informatics*, 2016.

[175] R. Zhang and O. Amft, "Monitoring chewing and eating in free-living using smart eyeglasses," *IEEE J. Biomed. Heal. Informatics*, Apr. 2017.

[176] R. Zhang, S. Bernhart, and O. Amft, "Diet eyeglasses: Recognising food chewing using EMG and smart eyeglasses," in *2016 IEEE 13th Int. Conf. Wearable Implant. Body Sens. Networks*. IEEE, Jun. 2016, pp. 7–12.

[177] R. Zhang, M. Freund, O. Amft, J. Cheng, B. Zhou, P. Lukowicz, S. Fernando, and P. Chabrecek, "A generic sensor fabric for multi-modal swallowing sensing in regular upper-body shirts," in *Proc. 2016 ACM Int. Symp. Wearable Comput. - ISWC '16*. New York, New York, USA: ACM Press, 2016, pp. 46–47.

[178] E. Thomaz, I. Essa, and G. D. Abowd, "A practical approach for recognizing eating moments with wrist-mounted inertial sensing," in *Proc. 2015 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. - UbiComp '15*. New York, New York, USA: ACM Press, 2015, pp. 1029–1040.

[179] World Health Organization, "Preparation and use of food-based dietary guidelines," Joint FAO/WHO consultation, Tech. Rep., 1998.

[180] N. Slimani, *et al.*, "The EPIC nutrient database project (ENDB): a first attempt to standardize nutrient databases across the 10 European countries participating in the EPIC study," *Eur. J. Clin. Nutr.*, vol. 61, no. 9, pp. 1037–1056, Sep. 2007.

[181] C. Woolhead, M. J. Gibney, M. C. Walsh, L. Brennan, and E. R. Gibney, "A generic coding approach for the examination of meal patterns." *Am. J. Clin. Nutr.*, vol. 102, no. 2, pp. 316–23, Aug. 2015.

[182] Public Health England, "National Diet and Nutrition Survey. Results from Years 1-4 (combined) of the Rolling Programme (2008/2009 - 2011/2012) Appendix R," Public Health England, Tech. Rep., 2014.

[183] Irish Universities Nutrition Alliance, "National Adult Nutrition Survey Summary Report," Irish Universities Nutrition Alliance, Cork, Ireland, Tech. Rep., 2011.

[184] R. M. Leech, A. Worsley, A. Timperio, and S. A. McNaughton, "Understanding meal patterns: definitions, methodology and impact on nutrient intake and diet quality," *Nutr. Res. Rev.*, vol. 28, no. 01, pp. 1–21, Mar. 2015.

[185] A. P. Hearty and M. J. Gibney, "Analysis of meal patterns with the use of supervised data mining techniques–artificial neural networks and decision trees." *Am. J. Clin. Nutr.*, vol. 88, no. 6, pp. 1632–42, Dec. 2008.

[186] H. B. Forslund, A. K. Lindroos, L. Sjöström, and L. Lissner, "Meal patterns and obesity in Swedish women - a simple instrument describing usual meal types, frequency and temporal distribution," *Eur. J. Clin. Nutr.*, vol. 56, no. 8, pp. 740–747, Jul. 2002.

[187] D. Gregori, F. Foltran, M. Ghidina, and P. Berchialla, "Understanding the influence of the snack definition on the association between snacking and obesity: a review," *Int. J. Food Sci. Nutr.*, vol. 62, no. 3, pp. 270–275, May 2011.

[188] K. Murakami and M. B. E. Livingstone, "Eating Frequency Is Positively Associated with Overweight and Central Obesity in US Adults," *J. Nutr.*, vol. 145, no. 12, pp. 2715–2724, Dec. 2015.

[189] K. Adolphus, N. Bellissimo, C. L. Lawton, N. A. Ford, T. M. Rains, J. Totosy de Zepetnek, and L. Dye, "Methodological Challenges in Studies Examining the Effects of Breakfast on Cognitive Performance and Appetite in Children and Adolescents," *Adv. Nutr. An Int. Rev. J.*, vol. 8, no. 1, pp. 184S–196S, Jan. 2017.

[190] R. M. Leech, A. Worsley, A. Timperio, and S. A. McNaughton, "Characterizing eating patterns: a comparison of eating occasion definitions," *Am. J. Clin. Nutr.*, vol. 102, no. 5, pp. 1229–1237, Nov. 2015.

[191] J. M. Hess, S. S. Jonnalagadda, and J. L. Slavin, "What Is a Snack, Why Do We Snack, and How Can We Choose Better Snacks? A Review of the Definitions of Snacking, Motivations to Snack, Contributions to Dietary Intake, and Recommendations for Improvement," *Adv. Nutr. An Int. Rev. J.*, vol. 7, no. 3, pp. 466–475, May 2016.

[192] G. H. Johnson and G. H. Anderson, "Snacking Definitions: Impact on Interpretation of the Literature and Dietary Recommendations," *Crit. Rev. Food Sci. Nutr.*, vol. 50, no. 9, pp. 848–871, Sep. 2010.

[193] A. Warde and L. Yates, "Understanding Eating Events: Snacks and Meal Patterns in Great Britain," *Cult. Soc.*, vol. 20, no. 1, 2017.

[194] A. Chamontin, G. Pretzer, and D. A. Booth, "Ambiguity of 'snack' in British usage," *Appetite*, vol. 41, no. 1, pp. 21–29, Aug. 2003.

[195] M.-P. St-Onge, J. Ard, M. L. Baskin, S. E. Chiuve, H. M. Johnson, P. Kris-Etherton, and K. Varady, "Meal Timing and Frequency: Implications for Cardiovascular Disease Prevention: A Scientific Statement From the American Heart Association," *Circulation*, vol. 135, no. 9, 2017.

[196] L. Yates and A. Warde, "The evolving content of meals in Great Britain. Results of a survey in 2012 in comparison with the 1950s," *Appetite*, vol. 84, pp. 299–308, Jan. 2015.

[197] C. E. O'Neil, C. Byrd-Bredbenner, D. Hayes, L. Jana, S. E. Klinger, and S. Stephenson-Martin, "The Role of Breakfast in Health: Definition and Criteria for a Quality Breakfast," *J. Acad. Nutr. Diet.*, vol. 114, no. 12, pp. S8–S26, Dec. 2014.

[198] B. Wansink, C. R. Payne, and M. Shimizu, ""Is this a meal or snack?" Situational cues that drive perceptions," *Appetite*, vol. 54, no. 1, pp. 214–216, Feb. 2010.

[199] J. M. Marx, D. A. Hoffmann, and D. R. Musher-Eizenman, "Meals and snacks: Children's characterizations of food and eating cues," *Appetite*, vol. 97, pp. 1–7, 2016.

[200] N. A. Younginer, C. E. Blake, K. K. Davison, R. E. Blaine, C. Ganter, A. Orloski, and J. O. Fisher, ""What do you think of when I say the word 'snack'?" Towards a cohesive definition among low-income caregivers of preschool-age children," *Appetite*, vol. 98, pp. 35–40, Mar. 2016.

[201] M. Lennernas and I. Andersson, "Food-based Classification of Eating Episodes (FBCE)," *Appetite*, vol. 32, no. 1, pp. 53–65, Feb. 1999.

[202] A. L. Olea López and L. Johnson, "Associations between Restrained Eating and the Size and Frequency of Overall Intake, Meal, Snack and Drink Occasions in the UK Adult National Diet and Nutrition Survey," *PLoS One*, vol. 11, no. 5, p. e0156320, May 2016.

[203] R. de Oliveira Santos, R. M. Fisberg, D. M. Marchioni, and V. Troncoso Baltar, "Dietary patterns for meals of Brazilian adults," *Br. J. Nutr.*, vol. 114, no. 05, pp. 822–828, Sep. 2015.

[204] S. L. Johnson, R. E. Boles, and K. S. Burger, "Using participant hedonic ratings of food images to construct data driven food groupings." *Appetite*, vol. 79, pp. 189–96, Aug. 2014.

[205] J. Riou, T. Lefèvre, I. Parizot, A. Lhuissier, and P. Chauvin, "Is There Still a French Eating Model? A Taxonomy of Eating Behaviors in Adults Living in the Paris Metropolitan Area in 2010," *PLoS One*, vol. 10, no. 3, p. e0119161, Mar. 2015.

[206] H. Leutheuser, D. Schuldhaus, and B. M. Eskofier, "Hierarchical, multi-sensor based classification of daily life activities: comparison with state-of-the-art al-

gorithms using a benchmark dataset." *PLoS One*, vol. 8, no. 10, p. e75196, Jan. 2013.

[207] V. Antila, J. Polet, A. Lamsa, and J. Liikka, "RoutineMaker: Towards End-User Automation of Daily Routines Using Smartphones," in *2012 IEEE Int. Conf. Pervasive Comput. Commun. Work.*, Lugano, 2012, pp. 399 – 402.

[208] D. Coyle and G. Doherty, "Clinical Evaluations and Collaborative Design : Developing new technologies for mental healthcare interventions," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Boston, MA, USA, 2009, pp. 2051–2060.

[209] K. Ellis, S. Godbole, J. Chen, S. Marshall, G. Lanckriet, and J. Kerr, "Physical activity recognition in free-living from body-worn sensors," in *Proc. 4th Int. SenseCam Pervasive Imaging Conf. - SenseCam '13*. New York, New York, USA: ACM Press, Nov. 2013, pp. 88–89.

[210] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A Public Domain Dataset for Human Activity Recognition Using Smartphones," in *2013 Eur. Symp. Aritifical Neural Networks, Comput. Intell. Mach. Learn.*, Bruges, Belgium, 2013, pp. 437 – 442.

[211] K. Zhai and J. L. Boyd-graber, "Online Latent Dirichlet Allocation with Infinite Vocabulary," in *Proc. 30th Int. Conf. Mach. Learn.*, 2013, pp. 561–569.

[212] I. Žliobaitė, "Learning under Concept Drift: an Overview," Faculty of Mathematics and Informatics, Vilnius University, Vilnius, Lithuania, Tech. Rep., 2009.

[213] S. M. Gerrish and D. M. Blei, "A Language-based Approach to Measuring Scholarly Impact," in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, 2010, pp. 375 – 382.

[214] D. M. Blei and S. M. Gerrish, "Dynamic Topic Models and the Document Influence Model C++," 2011. [Online]. Available: https://code.google.com/archive/p/princeton-statistical-learning/downloads

[215] M. A. G. Silva, "DTM executable," 2014. [Online]. Available: https://github.com/magsilva/dtm

[216] R. Rehurek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in *Proc. Lr. 2010 Work. New Challenges NLP Fram.* Valletta, Malta: ELRA, May 2010, pp. 45–50.

[217] J. Synnott, L. Chen, C. Nugent, and G. Moore, "IE Sim – A Flexible Tool for the Simulation of Data Generated within Intelligent Environments," in *Int. Jt. Conf. Ambient Intell.*, Nov. 2012, pp. 373–378.

[218] C. Krzyska, "Smart House Simulation Tool," Master's thesis, Technical University of Denmark, 2006.

[219] T. V. Nguyen, J. G. Kim, and D. Choi, "ISS: The Interactive Smart home Simulator," in *2009 11th Int. Conf. Adv. Commun. Technol.*, vol. 3, Feb. 2009, pp. 1828–1833.

[220] S. Helal, J. W. Lee, S. Hossain, E. Kim, H. Hagras, and D. Cook, "Persim - Simulator for Human Activities in Pervasive Spaces," in *2011 Seventh Int. Conf. Intell. Environ.*, Jul. 2011, pp. 192–199.

[221] B. Kormanyos and B. Pataki, "Home Sensor Simulator Software," 2013. [Online]. Available: http://home.mit.bme.hu/{~}kormi/simulator/

[222] ——, "Multilevel simulation of daily activities: Why and how?" in *2013 IEEE Int. Conf. Comput. Intell. Virtual Environ. Meas. Syst. Appl.*, Jul. 2013, pp. 1–6.

[223] K. Hoffmann, M. B. Schulze, A. Schienkiewitz, U. Nöthlings, and H. Boeing, "Application of a new statistical method to derive dietary patterns in nutritional epidemiology." *Am. J. Epidemiol.*, vol. 159, no. 10, pp. 935–44, May 2004.

[224] Public Health England, "National Diet and Nutrition Survey. Results from Years 1-4 (combined) of the Rolling Programme (2008/2009 – 2011/12)," Public Health England, Tech. Rep., 2014.

[225] NatCen Social Research, MRC Human Nutrition Research, and University College London Medical School, *National Diet and Nutrition Survey Years 1-4 2008/09 - 2011/12*, 7th ed. Colchester, Essex: UK Data Archive, 2015.

[226] Public Health England, "National Diet and Nutrition Survey. Results from Years 1-4 (combined) of the Rolling Programme (2008/2009 – 2011/12) Appendix A," Public Health England, Tech. Rep., 2014.

[227] NatCen Social Research, MRC Human Nutrition Research, and University College London Medical School, "National Diet and Nutrition Survey Rolling Programme Years 1-4 2008/2009-2011/2012 List of Variables for UK Core Sample Data," Public Health England, Tech. Rep., 2014.

[228] H. Syrad, C. H. Llewellyn, L. Johnson, D. Boniface, S. A. Jebb, C. H. M. van Jaarsveld, and J. Wardle, "Meal size is a critical driver of weight gain in early childhood." *Sci. Rep.*, vol. 6, no. 28368, Jun. 2016.

[229] L. Johnson, A. P. Mander, L. R. Jones, P. M. Emmett, and S. A. Jebb, "Energy-dense, low-fiber, high-fat dietary pattern is associated with increased fatness in childhood." *Am. J. Clin. Nutr.*, vol. 87, no. 4, pp. 846–54, Apr. 2008.

[230] V. Rosato, V. Edefonti, M. Parpinel, G. P. Milani, A. Mazzocchi, A. Decarli, C. Agostoni, and M. Ferraroni, "Energy Contribution and Nutrient Composition of Breakfast and Their Relations to Overweight in Free-living Individuals: A Systematic Review," *Adv. Nutr. An Int. Rev. J.*, vol. 7, no. 3, pp. 455–465, May 2016.

[231] B. W. Rothausen, J. Matthiessen, C. Hoppe, P. B. Brockhoff, L. F. Andersen, and I. Tetens, "Differences in Danish children's diet quality on weekdays v. weekend days." *Public Health Nutr.*, vol. 15, no. 9, pp. 1653–60, Sep. 2012.

[232] National Center for Health Statistics, "Introduction to National Health and Nutrition Examination Survey (NHANES)," U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Hyattsville, MD, Tech. Rep., 2015.

[233] Office for National Statistics. Social and Vital Statistics Division and Food Standards Agency, "National Diet and Nutrition Survey : Adults Aged 19 to 64 Years, 2000-2001," 2005. [Online]. Available: http://dx.doi.org/10.5255/UKDA-SN-5140-1

[234] National Diet and Nutrition Survey, "The National Diet & Nutrition Survey: adults aged 19 to 64 years —User Guide," UK National Data Archive, Tech. Rep., 2001.

[235] J. Seiter, W.-C. Chiu, M. Fritz, O. Amft, and G. Troster, "Joint segmentation and activity discovery using semantic and temporal priors," in *2015 IEEE Int. Conf. Pervasive Comput. Commun.* IEEE, Mar. 2015, pp. 71–78.