Analysing spatiotemporal sensor data on respiration, activity, and air pollution

Meeke Roet

Master of Science
Operational Research with Data Science
School of Mathematics
University of Edinburgh

2018

Abstract

In a pilot study as part of the Andhra Pradesh Children and Parents Study (APCAPS), based in India, data on respiration and physical activity was collected by 158 participants wearing a body-worn sensor called the *RESpeck*. This dissertation validates the value of this data for the overarching APCAPS project by showing that valuable and valid spatiotemporal attributes can be extracted. Secondly, it demonstrates that using an LSTM model the respiratory rate and trend can be predicted from the RESpeck signals with an MAE in the range of 1.5 - 2.7 and 0.4 - 1.75 BPM, respectively, for one to ten minutes ahead. This suggests that it may be possible to use the RESpeck for respiration monitoring in health settings to detect health threats such as sepsis or pneumonia. Finally, it is shown that air pollution affects the respiratory rate of asthma and COPD patients in the European TackSHS project, and that this could be helpful to improve respiratory rate predictions.

Acknowledgements

I would like to express my gratitude to my supervisor, Prof. D.K. Arvind, for his continuous, enthusiastic support and detailed feedback on my work, unchanged from before the project even started all the way to the end. I am also grateful to Darius Fischer for his work in preparing the data and his many helpful suggestions regarding both the technicalities of working with the data and the interpretation of the results.

I would furthermore like to thank all members of the research teams that provided me with the datasets that were central to the project: Sanjay Kinra, Santhi Bhogadi, Jessica Packer, and Kerri Wazny from APCAPS, and Sheila Keogan, Luke Clancy, and Shasha Li from TackSHS. Thanks also to Dr. Julian Hall, my second supervisor and personal tutor, for making it possible to take up a project in the School of Informatics and for his feedback.

Appreciation is also due to my fellow students from informatics for the discussions on our projects. Finally, many thanks to Amalie and Paula, who kept me company while writing, tried their best to get me into coffee, and patiently listened to and sometimes laughed at my complaints.

Like a machine learning algorithm, my output would not have been this good without all of your excellent inputs.

Table of Contents

List of abbreviations

AL	Activity level
APCAPS	Andhra Pradesh Children and Parents Study
BMI	Body mass index
BPM	Breaths per minute
CV	Cross-validation
LSTM	Long short-term memory model
MAE	Mean absolute error
REM	Rapid eye movement
(R)MSE	(Root) mean squared error
RNN	Recurrent neural network
RR	Respiratory rate
SWS	Slow-wave sleep
TS	Time series
TST	Total sleep time
WASF	Wake after sleep offset
WASO	Wake after sleep onset

Chapter 1

Introduction

The Andhra Pradesh Children and Parents Study (APCAPS) is a large, intergenerational cohort study based in 29 villages in southern India. It started as the follow-up of a study on childhood nutrition and has since collected extensive data on the cohort, ranging from socioeconomic characteristics to dietary habits and health (Kinra et al., 2014). The most recent effort of the APCAPS team is a pilot study in preparation for creating a 'sustainable longitudinal population study on the health effects of urbanisation by involving the community in research' (Research proposal APCAPS, 2018). This pilot study focused on using a body-worn sensor, the *RESpeck* (developed at the Centre for Speckled Computing in Edinburgh), to collect spatiotemporal data on respiration and physical activity of participants.

In order for the pilot data ($n = 158$) to help further the overarching APCAPS project, its value and validity have to be demonstrated. The first aim of this dissertation is therefore to analyse the pilot data in detail. More specifically, it will be shown what kind of information is contained in the data, not only explicitly, but also implicitly. To this extent, it is explored which additional features can be extracted from the data, summarising more information than what is measured directly by the RESpeck. Furthermore, the validity of the collected data is assessed by investigating if it displays similar patterns and relations as what has been established in the literature. Focal points for this will be personal characteristics (age, gender, etc.), sleeping patterns, respiration, and physical activity. Combined, this will validate the value of the RESpeck in the proposed study.

The second aim of this dissertation is to investigate whether it is possible to predict a person's respiratory rate based on the signals collected by the RESpeck. Together with heart rate, blood pressure, and temperature, the respiratory rate is one of the vital signs (McGee, 2012, p. 145). These vital signs are commonly monitored in patients, as abnormal deviations in them carry high diagnostic and prognostic value. For the respiration rate, a normal value at rest is around 20 breaths per minute (BPM). Depending on the situation, deviations roughly below 8 BPM or above 25 BPM are considered abnormal. This plays an important role in the detection of issues such as pneumonia (McGee, 2012, p. 146), sepsis (Ward and Levy, 2017, p. 9), and breathing problems after anaesthesia (Drummond et al., 2013). Timely and reliable detection of such issues is essential, because they come with great costs. Sepsis, for instance, is estimated to affect up to 31 million people globally every year, causing around 6 million deaths and a financial burden of US$ 24 billion (World Health Organization, 2017).

Nevertheless, to the author's best knowledge, there is no existing literature on predicting the respiratory rate ahead of time. Instead, research has focused on utilising the respiratory rate as an input to predict other outcomes directly (Várady et al., 2002; Segal et al., 2010; Argerich et al., 2016; Shanmathi and Jeyanthi, 2017). However, in some situations, such as during/after surgery or in the intensive care unit, healthcare practitioners may not be looking for specific diseases or events, but rather be interested in the development of the respiratory rate for monitoring purposes. Hence, predicting the respiratory rate itself is of value and will constitute an important part of this thesis. This will be done both directly on the respiratory rate, as well as on its trend. After estimating models on the APCAPS data, special attention will be paid to the relation between respiration and air pollution, and how this could be used to improve the respiratory rate predictions. Data for this was taken from the first eight subjects in another study, Tackling Secondhand Smoke (TackSHS), in which the subjects have a respiratory condition.

The results show, firstly, that valuable and valid spatiotemporal attributes can be extracted from the APCAPS pilot data. Secondly, using an LSTM model, the respiratory rate and trend can be predicted with an MAE in the range of 1.5 - 2.7 and 0.4 - 1.75 BPM, respectively, for one to ten minutes ahead. These predictions correspond reasonably well to the actual trend in the respiratory rate, which is a promising indicator that it may be possible to develop a system for respiration monitoring using the RESpeck. It is also shown that the respiratory rate in asthma and COPD patients is correlated to the air pollution exposure, and that it may be beneficial to include air pollution in models for respiratory rate prediction on patients with these respiratory conditions.

The rest of this report is structured as follows. The next chapter provides the background required for the project, firstly setting the framework of patterns and relations against which the APCAPS pilot data can be validated, and then introducing the main model used for predicting the respiratory rate, the LSTM model. After that, Chapter 3 will describe the APCAPS pilot data, how it was preprocessed, and which features were extracted from it. Chapter 4 continues by analysing relations in the data and comparing these to the patterns described in Chapter 2. Chapter 5 describes the implementation and results of the respiration prediction models. Chapter 6 builds on this by exploring the relation between air pollution and respiration, and whether this could be used to improve the predictions. Finally, Chapter 7 presents the conclusions of the work and suggestions for further research.

Chapter 2

Background

This chapter starts by giving an overview of the factors that influence sleep and physical activity patterns, which will be used to inform feature engineering and hypothesis forming later on. The corresponding sections are hence focused mainly on the factors about which information is contained in the dataset that will be used. Afterwards, the model that will be used for respiratory rate prediction is introduced.

2.1 Factors influencing sleep

2.1.1 Quantifying sleep quality

The first step in analysing sleep and the factors influencing it is to understand the concepts and measures used to quantify sleep quality. Although the nature of the data means that many of the measures will be unavailable in their exact definition during the analysis part of this thesis, a variety of them will be approximated. The details of this can be found in Chapter 3.

A good starting point is the guide of Shrivastava et al. (2014) on how to interpret the results of a sleep study. The first important definition is that of *total sleep time (TST)*, which is the total amount of time that a person was asleep, i.e. from sleep onset to sleep offset. This normally differs from the total time spent in bed, as people take some time to fall asleep and get up. The duration from when someone starts trying to fall asleep until actual sleep onset is called *sleep latency*. It is of interest in sleep studies, because minimal sleep latency indicates a lack of sleep, whereas high sleep latency is one of the symptoms of insomnia. Closely related is the concept of *sleep efficiency*: the time spent asleep expressed as a percentage of the total time in bed. Low sleep efficiency can be caused either by long sleep latency, or instead by a high number of waking minutes during the night. The latter is called *wake after sleep onset (WASO)*. Similarly, *wake after sleep offset (WASF)* is the time spent in bed after waking up. This is often high for people with depression and elderly, who may wake up early in the morning and cannot fall back asleep.

Sleep is characterised by four phases (Riley et al., 1985). Stage N1 is the first, in which a person is in very light sleep or a drowsy state. This is followed by stage N2 and N3, both deeper than their predecessor stage. Stage N3 is when events like sleep walking and sleep talking occur. It is also called slow-wave sleep (SWS) or delta sleep. The final stage is Rapid Eye Movement (REM) sleep, in which the body becomes almost completely paralysed. As a good night's sleep requires that the sleep time is distributed proportionally over the phases, sleep reports usually split up the TST by sleep phase (Shrivastava et al., 2014). This is called the *sleep architecture*. The time until REM sleep, or *REM latency*, is also often reported, as it is an indicator for several sleep-related disorders. Lastly, Shrivastava et al. mention *sleep fragmentation* as a common measure: the number of awakenings or shifts in sleep stage during the night. High sleep fragmentation can leave people tired even with a normal TST.

Although the above covers the measures most often encountered in the literature, a couple more could be identified. Several articles make use of subjective measures, for example self-judged sleep quality (Park et al., 2015), a feeling of sleepiness/exhaustion (Gonnissen et al., 2013; Ortega et al., 2010), and problems doing daily tasks (Karimi et al., 2016). Another measure is the occurrence of sleep disorder symptoms, such as apnoea-hypopnea events (Brower and Hall, 2001; Mendelson et al., 2016) or symptoms of Periodic Limb Movement Disorder (Brower and Hall, 2001). In some cases, the researchers use measures highly geared towards a specific application, e.g. sleep medication usage in Karimi et al. (2016). Finally, the use of questionnaires merging a variety of these indicators into one index is not uncommon. Examples are the PSQI questionnaire (Alfaris et al., 2015; Park et al., 2015) and Petersburg's sleep quality index (Karimi et al., 2016).

2.1.2 Personal factors

2.1.2.1 Age

The effect of age on sleep patterns has been described extensively in Morgan (1987). Although changes are perhaps most apparent in ageing children, this section will focus on adults and elderly to match the subjects whose data will be analysed. One clearly established relation is the low level of subjective sleep quality in the elderly. Morgan notes that older people report more frequent night and early morning awakenings, and rate their sleep as lighter. As it turns out, these three issues are not simply a consequence of an increasing inclination to complain with age, because they are confirmed by EEG scans. Although the time spent in bed increases with age, the time actually spent asleep decreases due to higher WASO. Moreover, the distribution of TST over the sleep stages changes towards more time in lighter sleep phases, matching the complaints about light sleep. Older people are also more likely to suffer from sleep breathing disorders (Subramanian et al., 2013). A final age-related change in sleep is an increase in naps taken during the day. Whether this is a cause for or consequence of the worsened quality of sleep at night remains unclear (Morgan, 1987).

2.1.2.2 Weight

Another factor affecting sleep is weight, in particular obesity. Gonnissen et al. (2013) make note of the concurrent increase in obesity and decrease in average self-reported sleep duration. Examining these two trends in more detail, they find an inverse association between sleep duration and BMI, as well as changes in the sleep architecture related to BMI changes. They hypothesise that these changes in sleep architecture cause more fragmented sleep, which on its turn leads to less self-control and increased appetite during the day, resulting in a higher BMI. They also posit physical activity as a potential third factor driving both BMI and sleep duration, since less active people have a higher BMI on average and tend to sleep shorter.

The inverse relation between BMI and sleep duration is confirmed in Gupta et al. (2002). Alfaris et al. (2015), who investigate the effect of weight loss on sleep quality and duration, only confirm it partly. They find that both sleep quality and duration increase after weight loss, but only in the short term. This may have to do with the fact that they look at the effects of a weight loss programme rather than general differences in weight.

In terms of sleep disturbances, there are certain sleep disorders that have been shown to be more common in the obese, such as sleep apnea and snoring (Morgan, 1987), which can lead to lighter sleep. However, Gupta et al. (2002) do not find that WASO is higher in obese people.

2.1.2.3 Gender

Like most biological processes, sleep knows some differences between men and women. First of all, men are more likely to suffer from sleep breathing disorders, most importantly snoring. At the same time, women are on average less satisfied with their sleep. Morgan (1987) suggests these two facts might be related, as snoring arguably is ‘more of a problem to one’s bed partner than it is to oneself’. He also reports that while in young age, men and women require the same amount of noise to be woken up, women become more sensitive to external noise than men as they age. Nevertheless, the majority of sleep indicators do not seem to depend on gender (Riley et al., 1985).

2.1.3 Lifestyle factors

2.1.3.1 Exercise

The effects of exercise on sleep can be considered either on a long-term or a short-term basis. The former is concerned with the differences in average sleep quality between people who exercise to different extents (‘between subject’). The latter focuses on the acute effects of exercising during the day on sleep quality at night (‘within subject’).

In general, exercise has a positive effect on sleep. In the long term, Mendelson et al. (2016) show that subjecting obese adolescents to a 12-week exercise programme improves sleep duration and sleep quality, as measured by sleep continuity and efficiency. Similar results were found by Karimi et al. (2016) for elderly men participating in an exercise programme and by Sherrill et al. (1998) for adults with sleep disorders after analysing their exercising habits. Awad et al. (2012) also show that exercise reduces sleep-disordered breathing, but they note that this seems to be caused for a large part by changes in the physique. Mendelson et al. and Karimi et al. do not control for this, but Sherrill et al. do.

The acute effects of exercise on sleep were reviewed in a meta-study by Youngstedt et al. (1996) including 38 studies on people without sleep disorders. They find that exercise significantly increases stage N2 sleep, SWS and TST, decreases REM sleep and REM latency, and does not have a significant effect on sleep latency and WASO. More recently, Roveda et al. (2011) confirmed that exercising in the morning improves TST and several other sleep quality measures in healthy men. This is contrasted by the results Mitchell et al. (2016), who do not observe a relation between moderate-vigorous physical activity, sedentary behaviour, and TST in women. A reason for this difference could be that Mitchell et al. use a wrist accelerometer to estimate physical activity and sedentary behaviour, whereas Roveda et al. had their subjects perform specific routines of strength and endurance exercises.

All in all, it can be concluded that the long-term effects of exercise have been established with more certainty than the short-term effects. Nevertheless, the consensus for both types of effects seems to be that they are at least mildy positive.

2.1.3.2 Alcohol consumption

Similar to exercise, the effects of alcohol consumption can be considered either with a long-term or short-term perspective. In Park et al. (2015), the long-term approach is taken. They show that higher scores on an alcohol-use disorder test are correlated to worse subjective sleep quality, shorter TST, more sleep disturbances, and more snoring in men, but not in women. They explain this by the fact that the study involved few women with high alcohol use. They do not find relations to sleep latency and sleep efficiency. Brower and Hall (2001) come to similar conclusions: alcoholics have lower TST and more disturbed sleep.

As for the short term, Ebrahim et al. (2013) provide a good review of studies on the acute effects of alcohol on sleep. Their conclusion is that no matter the amount, alcohol reduces sleep latency and the amount of disturbances in the first half of the night, but causes more disruption during the sleep later on. These effects are present across gender and age groups. The same conclusion can be found in Riley et al. (1985), who states that ‘by the morning [alcoholics] will have had less sleep than they would have without alcohol’.

2.1.3.3 Other

Riley et al. (1985) mention a few more influencing factors. The first one is the use of stimulants other than alcohol, which adversely affect sleep due to high concentrations of caffeine. Nicotine inhaled through smoking has similar effects. Napping can also influence sleep, but its effects differ from person to person.

One more factor, which is missing in Riley et al. as the problem did not exist in 1985, is the use of screens such as smartphones. It has been shown that longer screen time during the day and shortly before bed is related to poorer sleep quality, in particular shorter TST, lower sleep efficiency, and longer sleep latency (Christensen et al., 2016).

2.2 Factors influencing physical activity

2.2.1 Long-term

Seefeldt et al. (2002) give an overview of the factors affecting the average physical activity level of adults, as summarised below.

In terms of personal characteristics, a negative relation to physical activity levels is generally observed for increasing age, the presence of disabilities, low education, low socioeconomic status, and undernourishment. Females are also less active than males on average, especially in non-western cultures. The effect of socioeconomic status may be different in rural India, the source of this thesis's data, because a lower socioeconomic status in such a society often corresponds to physically demanding jobs.

Negative environmental influences stem mostly from geography and the social environment. Geographical location can be a barrier to exercise if the climate is unpleasant, neighbourhoods are unsafe, or exercise locations are rare. Socially speaking, certain cultures may look down on physical activity in general and on certain activity types (e.g. dancing), or on exercise for women. Such influences can be expected to play a role in the APCAPS dataset. Lastly, in contrast to what is often believed, physical activity levels during childhood are not systematically linked to those in adulthood.

2.2.2 Short-term

On a day-to-day basis, there are many factors that could influence a person's activity level: illness, social pressure, time availability, depletion (fatigue/soreness), weather, overeating, and so on. Not much research has been done on these relations, but still a large part of them were examined by Conroy et al. (2013) in a study on college students. They discovered that about one-third of the variability in daily physical activity is due to interpersonal differences, meaning that the rest is either caused by daily-changing factors or by random variation. They also find that people exercise more on weekdays and when they have more time available. The other tested factors, such as physical depletion, weather, and overeating, had no significant effect.

The acute effect of sleep on physical activity is one factor that stands out, because it did receive quite a bit of attention from researchers. However, the results are mixed. Mitchell et al. (2016) report no association between sleep quality (TST and sleep efficiency) and physical activity or sedentary behaviour the following day. On the other hand, Ortega et al. (2010) observe that feeling tired when awakening and short sleep are related to a lack of physical activity and excessive TV watching. In Pesonen et al. (2011), yet another conclusion is drawn, namely that low-quality sleep correlates to a *higher* activity level the following day in children. However, as children with a structural lack of sleep often display hyperactivity (Riley et al., 1985), that conclusion is likely not transferable to adults. Overall, the effect of sleep on physical activity is unclear.

2.3 Predicting respiration patterns

2.3.1 Methodology

As explained in the introduction, no studies about predicting the respiratory rate directly (rather than using it to predict other outcomes) have been carried out to the author's best knowledge. Therefore, the background work done in preparation for the respiratory rate predictions has focused on determining the appropriate methodology for this purpose.

As the respiratory rate is a continuously changing variable of which the past values will be used as inputs, its prediction falls in the domain of time series analysis. Traditional time series methods such as autoregressive and moving average models, however, may not be the most suitable for the task at hand for two reasons. Firstly, they cannot learn from multiple examples, so that it is not possible to estimate a general model from the recordings of multiple subjects. Secondly, they cannot (easily) be made nonlinear, while it is not unlikely that nonlinearities will be present.

To address these issues, machine learning methods were examined that fulfill these two requirements. Additionally, it was required that the method can handle multiple time series at once, because the sensors used for data collection also capture other, related variables besides the respiratory rate. The method should also be able to incorporate 'fixed' inputs alongside the time series, such as characteristics of the subject. Bontempi et al. (2013) describe ways to turn time series inputs into supervised learning problems, so that any machine learning technique can be applied to them. The downside of this approach is that the time series structure of the data is not exploited. Adding this requirement finally led to identification of the Long Short-Term Memory model (LSTM) as the best candidate for this purpose. LSTMs were first introduced in 1997 by Hochreiter and Schmidhuber and have since become popular for sequence forecasting due to their superior performance in a wide variety of fields (Bianchi et al., 2017), as well as increases in computing power that have made them more feasible to estimate. They are a particular type of recurrent neural network (RNN) that is able to learn dependencies over many periods (Hochreiter and Schmidhuber, 1997; Olah, 2015). Combined with the usual flexibility of neural networks in terms of inputs and outputs, this makes them suitable for respiratory rate prediction. They will be described in more detail in the next section.

Another important methodological choice is the selection of features to use. The lack of literature on respiratory rate prediction meant that this endeavour could not be inspired by previous, similar attempts. Nevertheless, some ideas for features were found in Prasertsung and Horanont (2016), who extracted several statistical characteristics from accelerometer data to help classify human activities. These are, for example, the maximum and minimum value, variance, and correlation between two time series, all in a certain time window. In the respiratory rate prediction problem, the inputs will not be the direct accelerometer data, but rather the derived respiratory rate and other attributes. Although this is data of a different type, the features used by Prasertsung and Horanont are valid for any time series. The choice of these and other features that were implemented is discussed in Chapter 5.

2.3.2 The LSTM model

In order to understand the LSTM model, it is necessary to first understand RNNs. This section therefore starts by explaining the workings of RNNs, largely following Bianchi et al. (2017), and then describes the modifications needed to arrive at an LSTM model.

A schematic overview of the RNN architecture is shown in Figure 2.1. Going from left to right, it shows three 'unfolded' steps of the recurring structure that forms the RNN, each step corresponding to one time step of the time series ($k-1$, k, $k+1$). The diagram also shows that the RNN has three layers, which are (from bottom to top) the *input layer*, *hidden layer*, and *output layer*. At each time step k, the network receives as input the time series value at that time, x_k. The x_k are thus scalars if the input consists of one time series, but they can also be vectors if there is more than one time series. Following Bianchi et al. (2017), the input dimension is denoted N_i.

The inputs are then multiplied by a weight matrix $W_i^h \in \mathbb{R}^{N_i \times N_h}$, where N_h is what is called the number of *units* in the hidden layer. The same happens to the output of the hidden layer from the previous time step, which is multiplied by a weight matrix $W_h^h \in \mathbb{R}^{N_h \times N_h}$. The two resulting vectors of length N_h are summed and processed by the hidden layer. More specifically, the units in the hidden layer apply a nonlinear transformation to the vector, such as a hyperbolic tangent or sigmoid function, represented by the polygon in the diagram. This results in the hidden layer output or *hidden state* h_k, also a vector of length N_h. The hidden state forms the 'memory' of the network, as it depends on past inputs and hidden states. It should be noted that while the diagram in Figure 2.1 contains only one hidden layer, it is possible to add more. In this case, each layer (except the first one) takes the output of the preceding layer as an input.

Finally, the output y_k is computed by multiplying the hidden state with another weight matrix, $W_h^o \in \mathbb{R}^{N_h \times N_o}$, where N_o is the desired output dimension. Optionally, it can also be transformed nonlinearly once more. The training of an RNN refers to finding the optimal weight matrices. This is often done by means of gradient descent algorithms, although a variety of other possibilities exist. More on the optimisation methods used in this dissertation can be found in Section 5.2.2.

The problem with regular RNNs as described above is that they are unable to learn long-term dependencies well. This is due to the vanishing gradient problem, where the gradients needed to optimise the neural network decrease or increase exponentially, leading to unstable results or an inability to solve the optimisation problem at all (Bengio et al., 1994). LSTMs solve this problem by introducing a so-called *memory unit* into the network (Hochreiter and Schmidhuber, 1997), which changes how the input and hidden state vectors are processed. An overview of the memory unit can be seen in Figure 2.2. It is a collection of five different nonlinear components, each with their own function.

The functions g_1 and g_2 are nonlinear and determine the *candidate hidden state* and *candidate output* in a similar fashion as in the normal RNN. Then there are three 'gates', indicated by a σ. The *forget gate*, σ_f, is a sigmoid function that combines the current input and the preceding output to determine how much of the preceding hidden state should be retained. The modified hidden state is then combined with the candidate hidden state, where the input gate, σ_g, determines how much of the candidate hidden state should flow into the new hidden state. The new hidden state is then processed by g_2 and the output gate, σ_o, which determines how much should be returned as output. Note that although the weight matrices seen in the ordinary RNN are not explicitly included in Figure 2.2, they are still present, just absorbed into the gates.

The reason that this solves the vanishing gradient problem is that no nonlinear transformations are applied directly to the hidden state anymore. In Figure 2.2, this shows in the fact that the line from h_{t-1} to h_t is interrupted only by multiplication and addition operators, which are both linear.

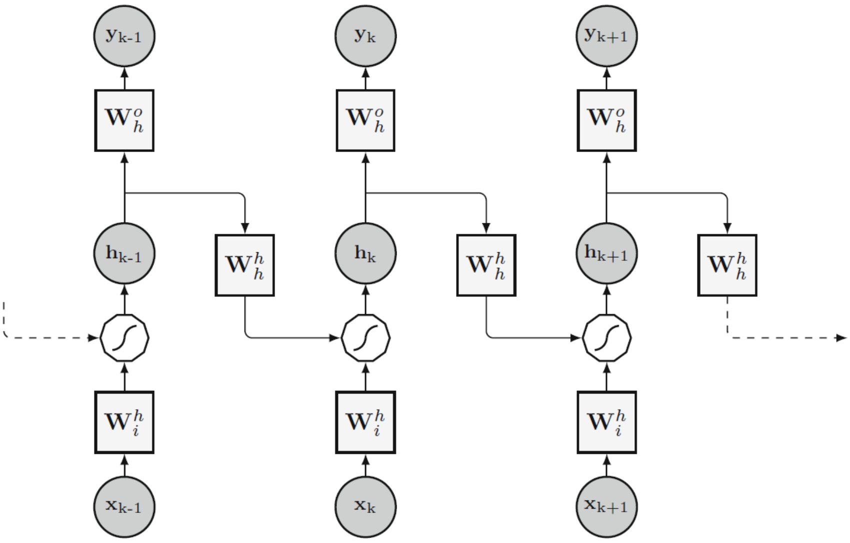


Figure 2.1: Diagram of the RNN architecture (Bianchi et al., 2017, p. 12).

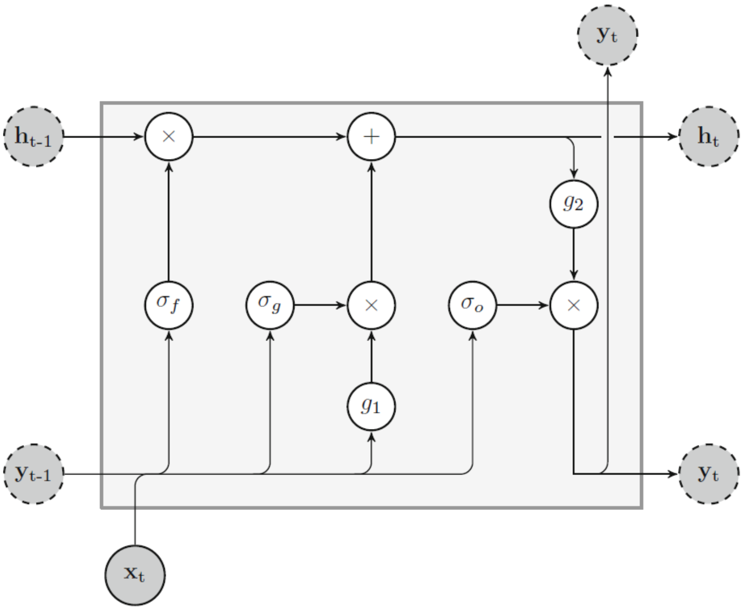


Figure 2.2: Diagram of an LSTM cell (Bianchi et al., 2017, p. 27).

Chapter 3

Data exploration, preprocessing and feature engineering

This chapter describes the data that was used, how it was preprocessed, and which features were extracted. It also details how outliers and other problematic data was dealt with. Finally, it sheds some initial light on relationships between attributes by examining their correlations.

3.1 Data description

The main dataset that is analysed in this thesis was collected as part of a pilot study under the aegis of the Andhra Pradesh Children and Parents Study (APCAPS), a large, intergenerational cohort study based near the city of Hyderabad in southern India. The pilot is intended as a first step towards creating a ‘sustainable longitudinal population study on health effects of urbanisation by involving the community in research’ (Research proposal APCAPS, 2018). Towards this purpose, subjects in the APCAPS cohort were invited to collect data for 1-3 days by wearing a *RESpeck*, a tiny, unobtrusive sensor that is taped to the abdomen (Figure 3.1). It records the linear acceleration of the abdomen along three axes, which is analysed to derive the respiratory rate (RR) and physical activity level (AL). For this thesis, data was available from a total of 158 subjects collected in two phases. Those with subject ID from 1 to 101 belonged to the first round of data collection, those with subject ID 201 to 257 to the second. The participants were 48% female. On average, they were aged 41 (standard deviation 14.4), had a resting pulse of 79 (11.7) and a BMI of 21.3 (3.6).

Figure 3.1: Participant wearing the RESpeck and holding a smartphone. Taken from Research proposal APCAPS (2018).

The RESpeck signals are logged at 12.5 Hz (12.5 times per second). To get more meaningful values and a more workable dataset, they were processed into attributes on a minutely basis by Darius Fischer at the Centre for Speckled Computing in Edinburgh. The resulting attributes are listed in the first five rows of Table 3.1. They are self-explanatory, except perhaps the following activity types: ‘movement’, denoting any type of movement that is not walking (e.g. riding a bike or bus), ‘wrong orientation’, denoting when the subject did not wear the sensor correctly, and ‘undetermined’, in case the activity type could not be determined with sufficient certainty. Due to noise in the RESpeck breathing signal when the wearer is moving a lot, the RR was missing for periods with high activity levels. To reduce the impact of this, the missing values were imputed by interpolating the values from the surrounding six minutes if there were at least two recordings in this window.

Besides the attributes created from the RESpeck signals, information about the participants’ location was derived from the phone they carried with them, which logged a GPS location approximately once every five seconds. As the RESpeck attributes are minutely, they were matched to the GPS coordinates that were recorded closest to the RESpeck timestamp. Furthermore, a tagged list of over 2000 locations was available from the APCAPS project, indicating for each location what kind of facility is located there (Table 3.2). A location from this list was added to each RESpeck observation by selecting the tagged location closest to the participant’s GPS location. If the closest tagged location was farther away than the GPS accuracy, it was left as a missing value. The proximity of two GPS locations was determined by calculating the distance

Table 3.1: Primary APCAPS variables (minutely).

Variable	Type	Description
Respiratory rate (RR)	Continuous	Average respiratory rate.
Standard deviation of respiratory rate	Continuous	Standard deviation of the respiratory rate measurements in the minute.
Activity level (AL)	Continuous	Average activity level, measured by the difference in acceleration compared to the previous measurement.
Activity type	Categorical	Inferred activity type, one from: sitting straight/standing, sitting bent forward, sitting bent backward, lying down on back, lying down facing right, lying down facing left, lying down on stomach, walking, movement, wrong orientation, undetermined.
Step count	Discrete	Number of steps taken.
GPS location	N/A	GPS coordinates of the subject's location.
Closest location	Categorical	Identifier of the closest tagged location.

d between them 'as the crow flies' using the Haversine formula:

$$\begin{aligned} a &= \sqrt{\sin^2\left(\frac{\phi_1-\phi_2}{2}\right)+\cos(\phi_1)\cdot\cos(\phi_2)\cdot\sin^2\left(\frac{\lambda_1-\lambda_2}{2}\right)} \\ c &= 2\cdot\arcsin\left(\sqrt{a}\right) \\ d &= c\cdot r \end{aligned} \tag{3.1}$$

where ϕ_i is the latitude coordinate of location i converted to radians, λ_i is the longitude coordinate of location i converted to radians, and $r = 6371000$ is the radius of the earth in meters. This completes the list of primary attributes in the dataset (Table 3.1).

3.2 Feature engineering

An overview of the secondary attributes, derived from the primary attributes, can be found in Table 3.3. They are explained in more detail below. All attributes were

Table 3.2: Location tag categories

Category	Subcategories
Shops for food/tobacco/alcohol	General store, restaurant/takeaway, other shop, meat/dairy shop, alcohol shop, street vendor (premade food), ration shop, street vendor (fresh produce), village market (fresh produce)
Physical activity site	Open space, gym, playground/sports (for children), open space (for children), playground/sports, swimming
Health service	Clinic/pharmacy, pharmacy, clinic, hospital, other health service, clinic (government), hospital (government), clinic/pharmacy (government)
Education service	Primary/middle school, preprimary school, middle/secondary school, other education service, primary/middle/secondary school, higher education, vocational college, middle/secondary/higher education, primary school, primary/higher education, higher education/vocational college

calculated both over the whole recording of a subject as well as on a day-to-day basis. This made a difference for the majority of the second-round and a few of the first-round subjects, because they recorded for more than 24 hours.

3.2.1 Averages and standard deviation by day/night

The RR and AL were averaged for each subject by day and by night, and the standard deviation was calculated. The minutely step count was also averaged, only for the day recordings. The reasoning behind splitting day and night variables is that the nightly RR is expected to be a ‘cleaner’ variable, because external influences such as exercise or pollution are most limited and (more or less) constant at night. This should make the analysis more reliable. At the same time, it should be noted that many of the participants only wore the sensor for approximately 24 hours, in which case there is at

Table 3.3: Secondary APCAPS variables. All variables are calculated over the total recording period as well as per recording day.

Variable	Type	Period	Description
Average RR, AL	Continuous	Day, night	Average respiratory rate, activity level.
Average step count	Continuous	Day	Average step count.
St. dev. RR, AL	Continuous	Day, night	Standard deviation of the respiratory rate, activity level.
Average walking distance	Continuous	Day	Average walking distance per minute.
Walking time fraction	Fractional	Day	Fraction of time spent walking.
Alcohol dummy	Binary	Day	Indicator for having visited an alcohol shop (1) or not (0).
Activity dummy	Binary	Day	Indicator for having visited a physical activity site (1) or not (0).
Health dummy	Binary	Day	Indicator for having visited a health service (1) or not (0).
Sleep interruptions	Discrete	Night	Number of times the subject got up during the night.
Wake after sleep onset (WASO)	Fractional	Night	Fraction of time spent not lying down during the night.
Turns	Discrete	Night	Number of times the subject switched sleep positions.
Napping dummy	Binary	Day	Indicator for having taken a nap (1) or not (0).

most one night worth of data per subject. This could introduce some noise, as there is no way to average out rare events such as a nightmare.

The daytime period was defined as any recordings between 8 a.m. and 8 p.m. and the nighttime as all minutes between 12 a.m. and 5 a.m., when the subject was lying down. Although a normal night's rest is longer for most people, the night window was chosen slightly tighter to reduce the chance of including recordings of lying down without being asleep.

3.2.2 Spatial and activity-related features

Several features were extracted from the GPS data. The first one is the average walking distance per minute during the day. This was determined by calculating the movement between observations according to the Haversine formula in Equation (3.1) and then averaging over minutes when the activity type was walking. For this purpose, the activity type 'walking' was subjected to two additional criteria, because without them, some impossible values occurred due to wrongly classified walking minutes (more on this observation in Section 3.3). Therefore, to be included as a walking minute, it was required that the movement in the minute was less than 200 meters, translating to 12 km/h. Although people can run faster than this, this is unlikely, so excluding these minutes for any walking statistics should not have a big effect. Secondly, it was ensured that the movement in the minute was not too large compared to the step count by requiring that the distance moved was at most 20% more than the step count, which can be interpreted as covering at most 1.2 meters per step. The reason behind this is that a normal step covers about 0.8 meters, so if someone is going more than 1.2 meters per step, it is highly unlikely that they are walking. The fraction of time spent walking during the day was also derived. Both these attributes are intended as alternative measures of physical activity besides the direct activity levels obtained through the RESpeck.

Secondly, it was calculated how much time the subjects spent in some of the location categories from the tagged location list. A threshold was set to count only periods in which the subject stayed in the same place for five minutes, to avoid counting locations that they were in reality just passing by. The (sub)categories hypothesised to be of interest were alcohol shops as a proxy for alcohol consumption, health services as a proxy for health, and physical activity sites as a proxy for exercise. Binary attributes were created for each of these categories, indicating if the subject visited a location of that type (1) or not (0).

Because the category of physical activity sites contains a few subcategories that do not necessarily imply exercise, namely open space, playground/sports for children, and open space for children (Table 3.2), the proxy for activeness was to be based only on the other three subcategories initially. However, it turned out that none of the participants spent at least ten minutes in such locations. The attribute was hence created for the overarching category of physical activity sites. Any results involving this attribute should therefore be interpreted with some caution.

The same holds for the alcohol shop and health service dummies. In the case of alcohol shops, visiting one does not necessarily imply drinking alcohol. After all, people can buy alcohol for later use or for others, or drink alcohol they already have stored at home without buying it the same day. For the health services, it is possible that the person bought medicines for someone else or went to the hospital as a visitor, rather than to be treated there themselves. Nevertheless, it will be interesting to investigate if the established effects of alcohol, exercise, and general health can be observed in the data through these attributes.

3.2.3 Sleep features

By wearing the RESpeck at night, subjects collected data during their sleep, albeit less than in a targeted sleep study. This data was used to derive several indicators of sleep quality, inspired by the standard measures identified in Section 2.1.1.

First of all, the number of sleep interruptions per night was calculated by counting how often the subject's activity type changed to something different than lying down during the night (excluding 'wrong orientation' and 'undetermined'). While this excludes awakenings without getting up and shifts in sleep stage, which are included in most common definitions of sleep fragmentation, it should still be a good proxy for sleep fragmentation. Furthermore, the number of minutes spent in activity types other than lying down was counted to approximate 'wake after sleep onset' (WASO). It was converted to a percentage of the total night recording to control for the fact that some participants only recorded part of the night. Although not following its strict definition, this attribute will be referred to as WASO.

Another indicator of sleep fragmentation is the number of turns during sleep, since a lot of turning could point towards restlessness and troubled sleep. This number was derived by counting how often the activity type switched between the four different positions of lying down. The most common measure of sleep, TST, is difficult to derive from the data, as there is no information on whether or not people were actually sleeping. Moreover, many of the subjects show an on-off pattern of lying down for a while in the evening, making it hard to pinpoint even just the bedtime, disregarding actual sleep. For these reasons, the TST was not implemented.

While the attributes above are all indicators of sleep quality, a binary attribute for napping was also created to serve as one of the potential influencing factors. Naps were detected by checking if someone lay down for at least 20 consecutive minutes during day time. Lastly, the standard deviation of the respiratory rate and activity level at night (Section 3.2.1) will be explored, following the reasoning that a high variance for these two attributes may be indicative of a night with many disruptions.

A note on this group of attributes is that they are always calculated on data between 12 a.m. and 5 a.m., which may be any part of a person's sleep depending on the bedtime. Some of the influences identified during the literature review, however, differed depending on the part of the night. For instance, alcohol consumption has a calming effect during the first half of the night, but leads to more sleep disruptions in the second. Due to the fixed time window defined as 'night', it is unlikely that effects this specific can be detected.

3.3 Outliers and other data issues

An overview of the issues that were identified in the data and how they were dealt with is given in Table 3.4. They are clarified in what follows.

Of the 158 participants, eight people did not record any data or their data was lost during transfer from the RESpeck to the phone or from the phone to the computer. For all other participants, the collected data was checked before the analysis for potential issues by examining summary reports such as the ones in Figure 3.2 and 3.3, as well as interactive maps showing the subject's location and other information per minute, a screenshot of which is shown in Figure 3.4. The summary reports give an overview of the activity types, activity levels, and respiratory rate over the recording period. Figure

Table 3.4: Overview of problematic subjects and solutions

Problem	Affected subjects	Solution
Defective sensor	78, 85, 90, 92, 97, 99, 101	Exclude
No data recorded or data lost	204, 205, 206, 208, 210, 226, 230, 236	Exclude
Sensor worn incorrectly or left somewhere (whole recording)	80, 203, 207, 215, 216, 227, 234, 245, 253	Exclude
Sensor worn incorrectly, left somewhere or disconnection issues (for parts of recording)	3, 48, 81, 202, 214, 221, 228, 231, 232, 240, 242, 246, 250, 251, 256	Exclude problematic parts
Recording less than two hours	39, 59, 233, 247	Exclude
Unreliable walking distance (> 3km/h)	24, 63	Impose logical conditions on walking classification
Unreliable WASO (> 0.3)	13, 20, 221, 246	Set affected variables as missing
Too many sleep interruptions due to misclassification of lying down vs. sitting leaning backward	64	Set affected variables as missing

3.2 is an exemplary recording, showing a clear day and night period and no gaps in the recording. At night, the activity level and respiratory rate are both low and relatively constant.

Figure 3.3 on the other hand is an example of a problematic set of data. The report shows characteristics of two issues that multiple subjects were affected by. First of all, some subjects left the RESpeck lying somewhere instead of wearing it, which shows up in the report as long stretches of lying down paired with a high respiratory rate. Secondly, the RESpeck sometimes disconnected from the phone, leaving it unable to store the recordings. In the figure, these are the interrupted stretches of activity type and activity level. The way the RESpeck works is that the last stored recording is spread out over the disconnection period, which skewed some of the features that were calculated.

A third issue that is not present in Figure 3.3 but occurred in a few other participants was incorrect wearing of the sensor, more specifically upside down, rendering the corresponding minutes useless. Fourthly, seven of the first-round participants used a defective RESpeck sensor that invalidated their recordings, an issue that was detected due to people having very high step counts while 'lying down'. The defective RESpeck was removed for the second round of data collection and the affected subjects were excluded from the analysis. Lastly, four subjects who recorded less than two hours in total were removed, as such short recordings are not representative of a daily cycle. These issues combined led to the exclusion of complete recordings for 28 subjects and parts for another 15.

After examining the subjects' summary reports, histograms were created of all continuous features to spot further irregularities. The first observation was the very high

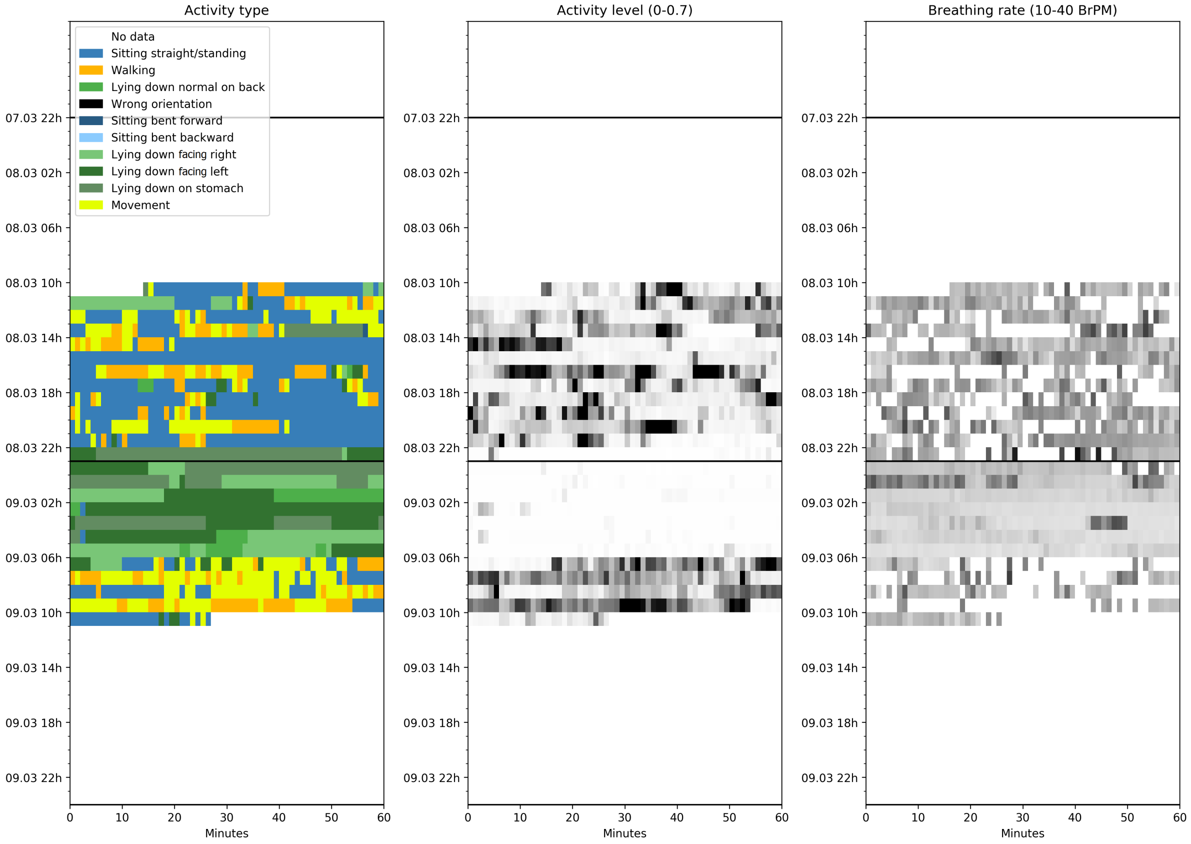


Figure 3.2: Summary report of exemplary collection of data, subject 24 (plot created by Darius Fischer).

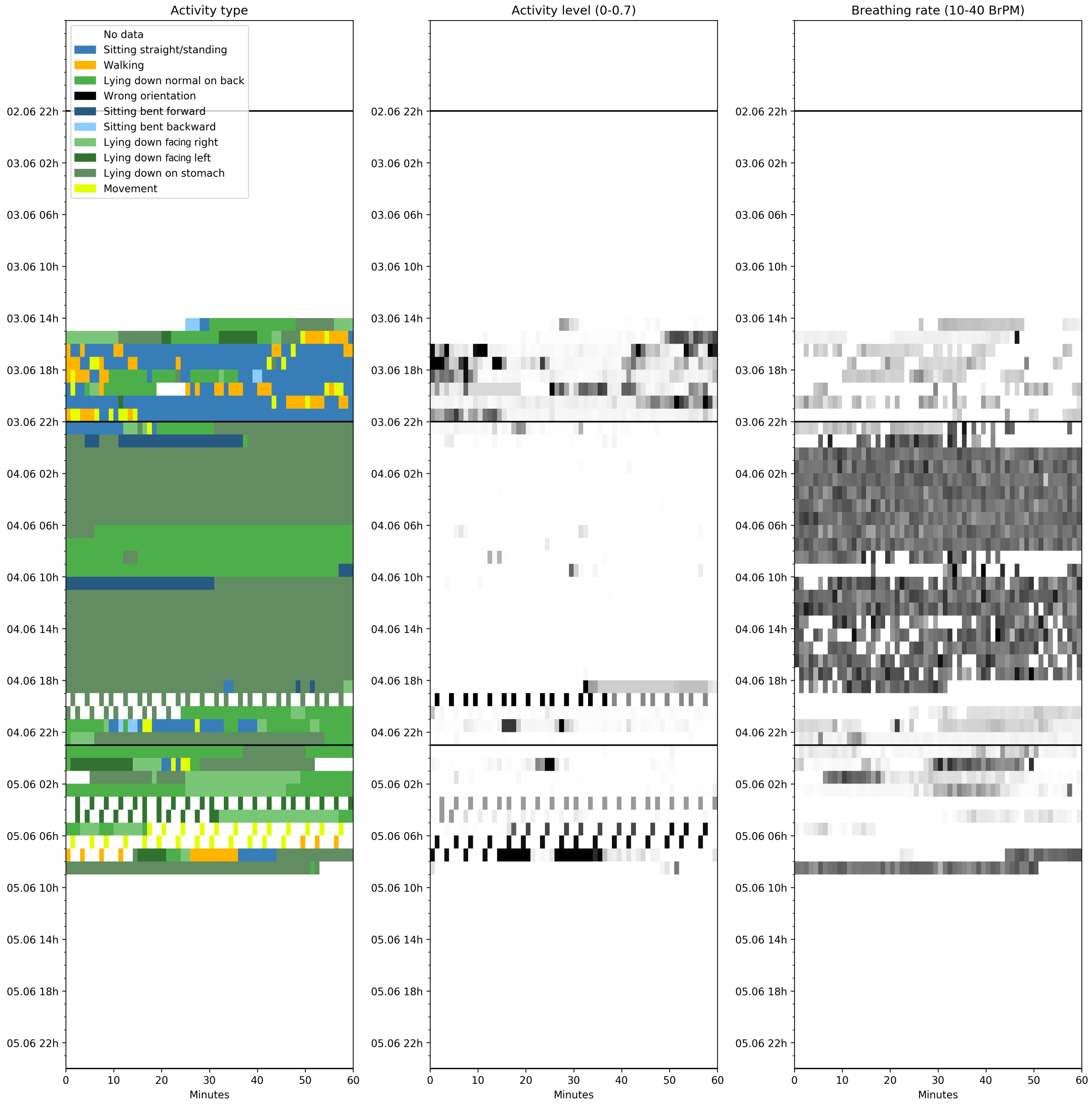
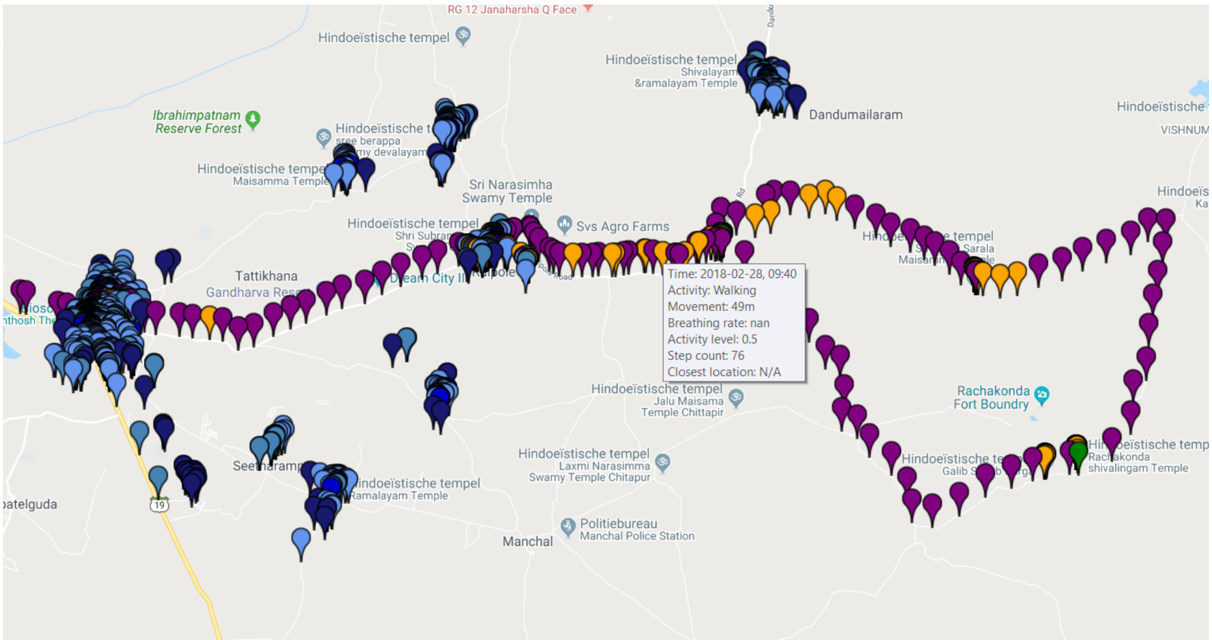


Figure 3.3: Summary report of problematic collection of data, subject 207 (plot created by Darius Fischer).

values of the average walking distance per minute. Several subjects had values of over 40 meters per minute (2.4 km/h). As this attribute was calculated by averaging the walked distances over the total number of minutes recorded, a speed of 2.4 km/h implies that a subject has walked at a normal walking pace of 5 km/h for approximately half the day. A look at the summary reports of the subjects in question showed that in reality, the time spent walking was much less than that for all of them. Further investigation showed that some of the minutes classified as walking had way too high



Time: 2018-02-28, 09:40
Activity: Walking
Movement: 49m
Breathing rate: nan
Activity level: 0.5
Step count: 76
Closest location: N/A

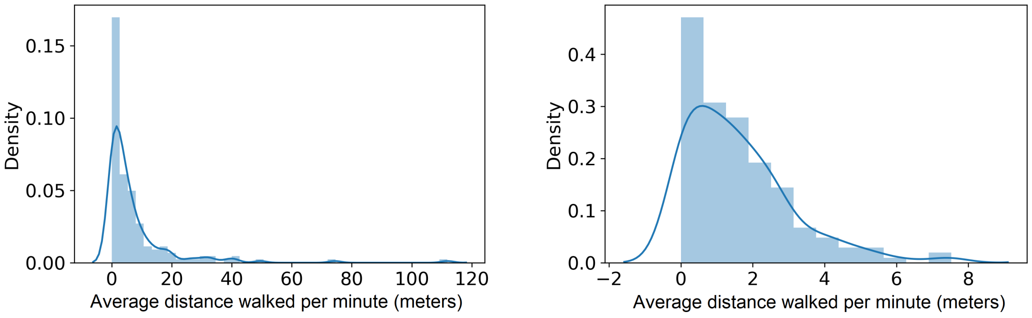
Figure 3.4: Sample map with movement and tagged locations, subject 2. Purple = movement, orange = walking, green = sitting/standing, blue (shades) = location categories.

movement for anyone to be walking. This led to the conclusion that, despite the existence of the activity type category ‘movement’, the activity classification algorithm had some troubles distinguishing walking from other activities, such as riding a bike, motorbike or bus. This was solved by imposing additional distance and step count criteria on the walking classification, as explained in Section 3.2.2. Figure 3.5 shows that this took care of the outliers effectively.

Two other attributes with outliers were spotted, all of which were caused either by RESpecks disconnecting from the phone or confusion between lying down and sitting down leaning backward, judging from the summary reports. The attributes in question are WASO (Figure 3.6), with some people awake almost half the night, and sleep interruptions (Figure 3.7), with one person waking up 15 times in a night. These values were set to be missing for the affected subjects.

3.4 Basic analysis of APCAPS dataset

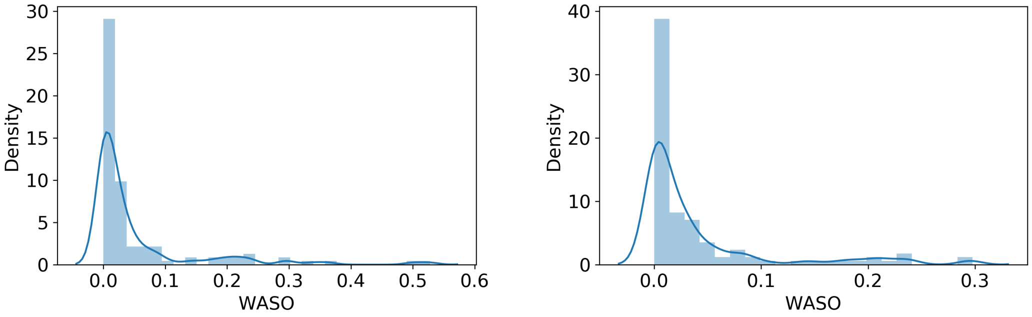
To get an initial feel for the data, Figure 3.8 shows a correlation matrix of the main attributes in the dataset related to RR and AL, as well as their lagged values up to



(a) Without extra conditions on walking.

(b) With extra conditions on walking.

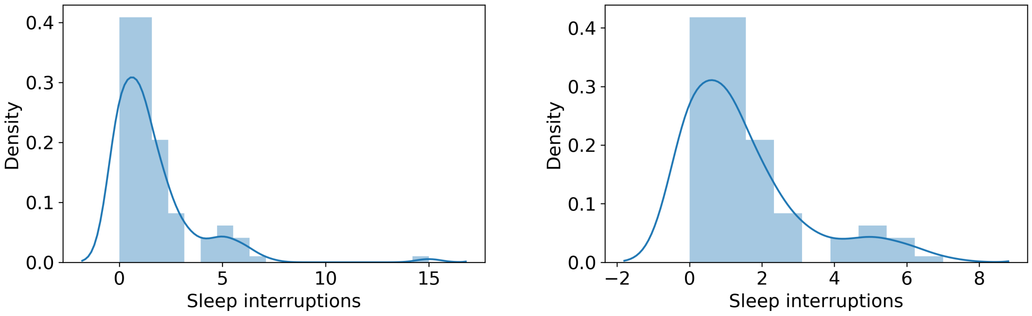
Figure 3.5: Histogram of average walking distance per minute.



(a) Before outlier removal.

(b) After outlier removal.

Figure 3.6: Histogram of WASO.



(a) Before outlier removal.

(b) After outlier removal.

Figure 3.7: Histogram of sleep interruptions.

three minutes ago. It can be seen that both the RR and AL exhibit a high level of autocorrelation (around 0.7) which decreases with lag time. Furthermore, the AL is highly correlated to the step count (0.82), which is in line with expectations. The RR and AL also have a mild correlation to each other of slightly over 0.20. This correlation fades only very slowly as the AL is taken with a higher lag, suggesting that changes in the AL have effects that last at least a few minutes. However, not only is the lagged AL correlated to the current RR, the reverse is also true. This is counterintuitive, as there is no reason to assume previous breathing changes are related to current changes in activity. The observed correlation could have to do with the fact that the RR is quite strongly correlated to itself.

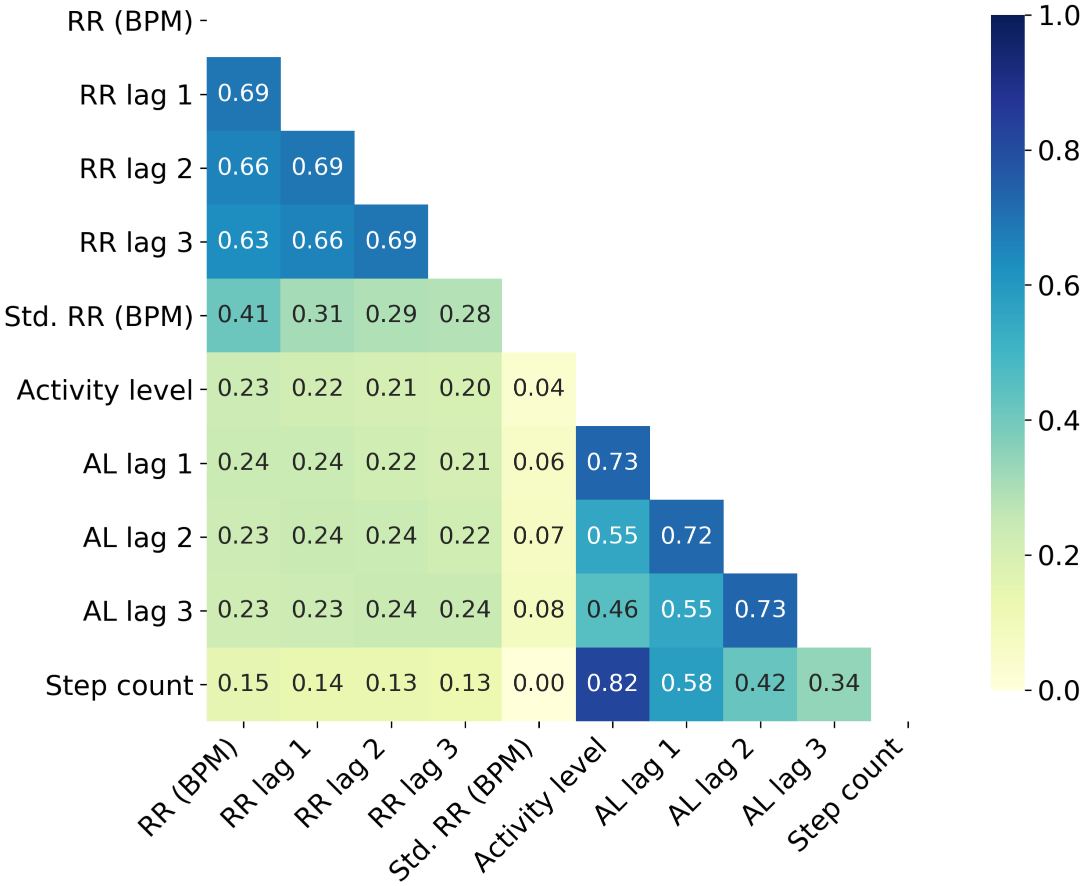


Figure 3.8: Matrix of correlations averaged over the subjects.

Chapter 4

Analysis of the APCAPS dataset

This chapter reports on the analysis of the APCAPS dataset and in the process validates the data collection and analysis methods against published results. This is managed at two levels: assessing the differences between the subgroups, and analysing the relationship between the different attributes.

4.1 Differences between subgroups

Four attributes of binary characteristics were created to relate the subjects' activities over the 24-hour period, such as visits to physical activity sites, alcohol shops, and health service providers, and whether or not they took a nap during the day. A fifth attribute was created to indicate waking up during the night, should the subject have at least one instance of sleep interruptions. The gender of the subjects was the sixth distinction. The differences between the subgroups defined by these six attributes were assessed by conducting statistical tests on the subgroup means and distributions.

4.1.1 Test procedure

When testing the difference between two means, it is important to know if the two samples are independent of each other or paired, meaning that for each observation in group 1, there is a matching observation in group 2, for example two test scores from the same person. It is clear that the subgroups defined by the binary attributes are not paired, as they can have different sizes and be based on recordings by different people. However, they are not 100% independent either. This will be explained below.

The desired testing can be done either on a subject-to-subject basis, treating one subject as one observation, or on a day-to-day basis, treating each recorded day/night as one observation, regardless of the subject. In this application, the latter makes more sense, because for instance, having visited an activity site once during three recording days is unlikely to be related to the average RR rate during those days. On the other hand, we may expect to see an effect when comparing the average RR on days where an activity site is visited to days when there is not. This means that the day-to-day variation is what is of interest[1]. This way of grouping implies that the subgroups are not completely independent, as multiple days recorded by the same participant may end up in different subgroups. Moreover, if the people who recorded multiple days differ substantially in some respect from the people who recorded only one day, then the results will be biased towards the people with longer recordings.

Despite these two concerns, it is assumed here that the subgroups are independent for two reasons. The main factor determining the recording time for the participants is whether they were in the first or the second round of data collection, because people in the first round were asked to record for at least one day, whereas in the second round they were encouraged to record for three days. As the subjects for the first or second round were selected randomly, there is no reason to assume that substantial differences between long and short recordings are present. Furthermore, independent tests are generally more conservative than tests controlling for dependencies, so in effect, assuming independency will lead to underestimation of effects.

The most common test for the difference between two means is the two-sample t-test. There are three assumptions to this test:

1. The two samples should be randomly drawn independently of each other.
2. The populations that the samples are drawn from should be normally distributed.
3. The populations that the samples are drawn from should have equal variances.

As explained, the first assumption is assumed to hold. The second assumption was assessed per subgroup by means of the normality test based on skewness and kurtosis developed in D'Agostino (1971) and D'Agostino and Pearson (1973) and implemented

[1] The exception to this are the tests involving solely personal characteristics: age, BMI, and pulse between men/women.

in the python library *SciPy* (SciPy, 2018). The test indicated that normality could be assumed to hold for the BMI, pulse, and average RR day/night for all subgroups. As there is no reason why the variances of the subgroups should be different for those attributes, t-tests were used for these comparisons. The test statistic is calculated as:

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{4.1}$$

where μ_i is the mean of sample i, σ^2 is the variance, and n_i is the number of observations in sample i. The statistic is compared to the t-distribution with $n_1 + n_2 - 2$ degrees of freedom. The null hypothesis is that the means of the two groups are equal, the alternative hypothesis is that they are not.

It is not surprising that most of the other attributes are not normally distributed, as for example the sleep attributes are bound by zero on the left. For these comparisons, the Mann-Whitney U test was used, an approach also taken by Mendelson et al. (2016). The test is suitable as it does not require assumption 2 and 3 of the t-test. Another advantage is that it is less sensitive to small sample sizes, which is indeed an issue for some of the subgroups (activity site/alcohol shop attendance). The Mann-Whitney U test is a nonparametric test with a slightly different aim than the t-test, namely determining how likely it is that two samples come from the same population. The basic idea is to rank all observations from lowest to highest, examine to which sample the observations belong, and produce a test statistic based on the rankings per subgroup. The test statistic is calculated as the smallest value from:

$$u_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - r_1 \tag{4.2}$$

$$u_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - r_2 \tag{4.3}$$

where r_i is the sum of the ranks for the observations in sample i. It is compared to a table of critical values based on n_1 and n_2. The null hypothesis is that the samples come from the same population. The reason that the Mann-Whitney U test is not always used, even though it requires fewer assumptions than the t-test, is that the t-test has more power when its assumptions do hold.

4.1.2 Results

The results of the statistical tests are shown in Table 4.1 with significance indicated according to the appropriate test as explained[2]. The table also shows in parentheses the number of observations in both subgroups. There are blanks for the comparisons of sleep interruptions and WASO based on waking up or not, as the awakening attribute is derived from them.

A first observation is that there are no significant differences between the location-based subgroups, which is surprising given the rather strong evidence found in the literature that exercise and alcohol consumption have an effect on sleep quality for example. Other obvious relations, such as a higher average RR for people who visited an activity site, are also not observed to a significant extent. An explanation for this could be that the attributes are not precise enough in capturing the intended information about the subjects, i.e. exercise, alcohol consumption, and health status, a possibility already identified when the attributes were created. There are a few entries in Table 4.1 that suggest this, because they run counter to intuition. The group which visited an activity site, for example, was older on average and had a lower step count than those who did not, whereas younger people normally exercise more, and exercise should be positively related to the step count. At the same time, there are other differences that do make sense; people who visited an activity site woke up less often on average and people who visited an alcohol shop turned more in their sleep. This points towards another explanation for the lack of significance, namely that some of the subgroups are quite small, regularly containing under 20 observations for the activity- and alcohol-based subgroups. This makes it hard to detect effects.

The other three attributes do reveal some relationships. There are four highly significant ($p < 0.01$) differences between men and women, all of which are explainable. Firstly, the pulse of females is around 6 beats/minute higher than that of males. This is a known difference caused by the fact that women, including their hearts, are smaller on average than men, so that their heart has to beat more frequently to provide enough oxygen (Nio et al., 2015). The three attributes average daytime AL, average step count and the walking time fraction are each lower for women compared with men, matching the observations in literature that women are less active on average than men, especially so in non-western cultures (Seefeldt et al., 2002). The sleep attributes do

[2] Boxplots of the subgroups/attributes with significant differences are included in Appendix A.

Table 4.1: Differences between subgroup means (mean ‘yes’ minus mean ‘no’).

	Visited activity site	Visited alcohol shop	Visited health service	Took a nap	Woke up at night	Female
Age	2.68	−1.70	−0.98	−1.63	3.18	−0.14
	(22/144)	(15/151)	(38/128)	(117/49)	(84/40)	(66/68)
BMI	0.31	−1.10	0.80	0.19	−0.43	−0.49
	(22/144)	(15/151)	(38/128)	(117/49)	(84/40)	(66/68)
Pulse	−0.63	2.44	−3.13	1.14	−4.39*	5.75***
	(22/144)	(15/151)	(38/128)	(117/49)	(84/40)	(66/68)
Av. RR day	0.12	−0.01	−0.24	0.52*	0.21	0.36
	(22/144)	(15/151)	(38/128)	(117/49)	(75/36)	(82/84)
Av. RR night	0.69	0.14	−0.63	0.45	0.27	0.63
	(17/94)	(12/99)	(29/82)	(83/28)	(84/40)	(55/69)
Av. AL day	0.00	−0.01	0.02	−0.05***	0.01	−0.04***
	(22/144)	(15/151)	(38/128)	(117/49)	(75/36)	(82/84)
Av. AL night	0.00	0.00	0.00	0.01	0.00	0.00
	(17/94)	(12/99)	(29/82)	(83/28)	(84/40)	(55/69)
Std. RR day	−0.01	−0.16	−0.07	0.22***	0.11	0.14
	(22/144)	(15/151)	(38/128)	(117/49)	(75/36)	(82/84)
Std. RR night	−0.52	0.14	0.40	0.29	0.09	0.05
	(17/94)	(12/99)	(29/82)	(83/28)	(84/40)	(55/69)
Av. step count	−1.10	−1.09	0.43	−3.16**	0.31	−3.26***
	(22/144)	(15/151)	(38/128)	(117/49)	(75/36)	(82/84)
Walking time fraction	0.02	0.00	−0.01	−0.03	0.00	−0.01***
	(22/144)	(15/151)	(38/128)	(117/49)	(75/36)	(82/84)
Sleep interruptions	−0.49	−0.10	0.23	−0.54*		−0.20
	(16/94)	(12/98)	(29/81)	(82/28)		(55/68)
Turns	1.65	0.89	0.48	4.50	0.93	0.95
	(17/94)	(12/99)	(29/82)	(83/28)	(84/40)	(55/69)
WASO	0.01	0.01	−0.01	−0.01*		−0.01
	(17/90)	(12/95)	(28/79)	(81/26)		(52/68)

Number of observations in parentheses (yes/no).

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

not appear to be different between men and women, also matching the literature (Riley et al., 1985).

When splitting the subjects by napping behaviour, there are three (highly) significant and logical relations: the average AL of that day is lower, the standard deviation of the RR is higher, and the average step count is lower, all caused by the stretch of low ALs and RRs during the nap. There are also some mildly significant differences ($p < 0.10$), namely a higher average RR during the day, fewer sleep interruptions and lower WASO. Napping thus seems to improve sleep quality slightly, although its effects are said to differ from person to person in the literature review. Lastly, people who woke up at night had a lower pulse ($p < 0.10$). This is probably at least partly caused the fact that men made up 57% of the awakening observations as opposed to 52% of the non-awakening observations, and they have a lower pulse.

4.2 Linear regressions

Besides differences between subgroups, it is also interesting to see what influences certain continuous attributes. For this purpose, linear regression (OLS) was used. This allows for assessment of effect sizes after controlling for potential confounders. Dependent variables were the average RR during the day/night, the average AL during the day, and the sleep attributes. The average RRs during the day and night were regressed on the personal characteristics, average AL during the (preceding) day and the binary attributes for visiting activity sites and alcohol shops. The step count, average walking distance, and walking time fraction were excluded, because their high correlation to the average AL led to multicollinearity issues. As the most general measure summarising activity, the average AL was retained.

As the units of the average AL are not easily interpretable, this attribute was standardised by dividing with the standard deviation. As a consequence, its coefficients can be interpreted as the expected *ceteris paribus* change in the dependent variable when the average AL increases by one standard deviation. The other attributes were not standardised as their units make sense also without standardising. These coefficients should therefore be interpreted as the expected change in the dependent variable when the attribute in question changes by one unit, again, given that everything else remains the same.

The results of the two RR regressions are presented in Table 4.2. Most of the resulting coefficients, although not significant, are in line with expectations when the sign is considered. A higher BMI and average AL during the day are both related to a higher RR, and so is having visited an activity site. It is interesting to see this is also the case for the regression on the nightly RR, as it suggests that a higher RR caused by exercise/more activity during the day persists into the night. However, given the high standard errors, these tentative results need to be interpreted very carefully. The one attribute that is significant at the 5%-level is pulse for the average RR at night, indicating that people with a higher resting pulse also tend to have a higher 'resting' RR.

The third regression shown in the same table is for the average AL during the day. It includes the personal characteristics as controls and focuses on the sleep attributes as other explanatory variables, because literature indicated this might affect physical activity levels. The regression is based on much fewer observations than the other two (33 versus more than 100), as not many pairs of days and preceding nights were recorded; virtually all subjects who recorded only one day/night recorded the night after the day. Nevertheless, some effects are observed. Again, the pulse coefficient is significant at the 5%-level, but this time it shows a negative relation, i.e. people with a higher pulse are less active on average. This matches what is known about the resting heart rate of trained versus untrained people. Significant at the 10%-level is the positive coefficient for the number of turns during sleep. This seems counterintuitive when turning in sleep is taken as an indication of lower sleep quality, but it could be that a person's general level of activity influences both their AL when they are awake as well as when they are sleeping. Sleep interruptions and WASO are both not significantly related to the average AL on the following day, but it cannot be ruled out that this is due to small sample size.

The results for the sleep attributes are summarised in Table 4.3. The attributes most of interest here are age, weight, and average AL, as they are attributes that the literature indicated are related to sleep quality, but they were not assessed yet in Section 4.1. The other factors that did get asssessed there are still included as controls. It can be seen that in this dataset, BMI and average AL the preceding day are not significantly related to sleep quality. Age, on the other hand, shows strong results.

Sleep interruptions and WASO are significantly higher for older people ($p < 0.01$). Both coefficients seem small, but their meaning becomes clear with an example. Compared to someone who is 30 years younger, a 60-year old is expected to have approximately one more sleep interruption ($30 \cdot 0.03$) and be awake for nine more minutes each night ($30 \cdot 0.001 \cdot 300$, as the WASO is measured as a fraction of the period between 12 a.m. and 5 a.m., i.e. 300 minutes). With the night defined as only a 5-hour period in this analysis, the figures for a complete night will be even higher. The coefficient of age for number of turns is also significant, but negative. Although elderly often rate their sleep as worse, based on which one would expect a positive coefficient (more turning in sleep with older age), literature suggested that this has to do mainly with changes to the sleep stage profile. The negative coefficient can therefore be explained.

Table 4.2: Regression results for average RR and AL.

	Average RR day	Average RR night	Average AL day
const	19.450	12.127	4.848
	(1.583)	(2.760)	(1.700)
Age	0.005	−0.001	−0.023
	(0.010)	(0.016)	(0.014)
BMI	0.045	0.116	0.068
	(0.040)	(0.071)	(0.060)
Female	0.335	0.568	−0.066
	(0.285)	(0.475)	(0.363)
Pulse	0.020	0.053**	−0.033**
	(0.012)	(0.021)	(0.015)
Average AL day	2.903	0.486	
	(2.061)	(3.550)	
Activity dummy	0.095	0.684	
	(0.393)	(0.624)	
Alcohol dummy	0.046	0.321	
	(0.468)	(0.725)	
Sleep interruptions			−0.067
			(0.137)
Turns			0.034*
			(0.018)
WASO			0.446
			(4.624)
R^2	0.04	0.10	0.42
Observations	165	110	33

Standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 4.3: Regression results for sleep variables.

	Sleep interruptions	Turns	WASO
const	3.385	48.118	0.160
	(2.551)	(14.592)	(0.109)
Age	0.030***	−0.134**	0.001***
	(0.011)	(0.063)	(0.000)
BMI	−0.051	−0.235	−0.003
	(0.048)	(0.265)	(0.002)
Female	0.227	0.373	0.012
	(0.313)	(1.783)	(0.013)
Pulse	−0.021	−0.067	−0.001**
	(0.014)	(0.078)	(0.001)
Average RR day	−0.029	−0.803	−0.002
	(0.091)	(0.510)	(0.004)
Average AL day	1.677	−12.976	0.104
	(2.420)	(13.788)	(0.101)
Activity dummy	−0.452	2.098	0.009
	(0.416)	(2.336)	(0.017)
Alcohol dummy	−0.132	0.076	0.000
	(0.473)	(2.702)	(0.019)
Napping dummy	−0.166	3.514	0.000
	(0.375)	(2.122)	(0.016)
R^2	0.12	0.14	0.15
Observations	109	110	106

Standard errors in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Chapter 5

Respiratory rate prediction

This chapter describes the models developed for predicting the respiratory rate. It starts with a description of the baseline models, features, and target variables. It continues with an overview of the estimation strategy, more specifically the optimisation algorithm that was used and the tuning of hyperparameters, followed by a discussion of the results of all models.

5.1 Models

5.1.1 Baseline

Three baseline models were considered for different reasons. The first is taking the mean RR over the past *n* minutes (called the *n*-period mean from here on), as a very simple estimator which incorporates the fact that the RR is correlated over time. The second is continuing the slope over the past *n* minutes (called the *n*-period slope from here on), with the reasoning that the RR follows a trend that the slope may approximate.

The third baseline is a more sophisticated model based on linear regression (Ridge regression) on the past RR and AL. The choice was justified on similar basis to the *n*-period mean - it reflects that the RR is correlated over time, with the advantage that the coefficients for each lag can be different. As a result, the regression has the option of weighting recent RRs more heavily, and thereby mimicking the fade of correlation over time. A Ridge regression differs from normal linear regression in that an L2 penalty term is added to the objective function, changing it from:

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{5.1}$$

to

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{5.2}$$

with y the observed value, $\hat{y}$ the prediction, n the number of observations, β the coefficient, p the number of coefficients, and λ the penalty term. The advantage of adding this penalty term is to help prevent overfitting, which means that the model fits the training data 'too well', therefore generalising poorly to unseen data. The Ridge penalty term prevents overfitting by placing a penalty on the coefficient sizes, so that a trade-off is made between the goodness of fit as expressed by the residual sum of squares (RSS) and the size of the coefficients. This reduces overfitting, as the coefficients are shrunk to less than what would be ideal to fit the training data.

Another method following the same principle is Lasso regression, which uses L1 regularisation (i.e. penalties on the absolute value of coefficients) rather than L2. It has the property of creating sparsity, forcing some coefficients to zero, whereas Ridge regression shrinks all coefficients proportionally to the degree of contribution to the predictions. With this in mind, Ridge was chosen over Lasso, because there is no reason to assume that any particular lags of the RR are uncorrelated to the current RR and should get a coefficient of zero.

The choices of n for the n-period mean and slope, and lag structure and λ for the Ridge regression are detailed in Section 5.2.3. This practice of tuning the model's settings, also called *hyperparameters*, is called hyperparameter tuning.

5.1.2 Respiratory rate models

As explained in Section 2.3, the LSTM model was chosen for predicting the RR. Four versions of this model were estimated using different combinations of features. The models were estimated on the nighttime data only, because the RR during the day is confounded by too many factors that are not included in the dataset. The nighttime data is generally regarded as a measure of the resting RR and should therefore produce more meaningful results. This decision automatically led to exclusion of certain features that might actually correlate well with the RR, but are not useful when considering the night only, such as the activity type and location. Having noted this, the features making up the four models are described in the following sections.

5.1.2.1 Time series

The most basic version of the model uses the past RR and AL directly as two time series inputs, the logic being that the RR changes gradually over time and is furthermore influenced by the AL in the recent past, as the RR may stay elevated for a while after increased activity. Figure 5.1 shows the correlation between the current RR and the RR/AL from one to thirty minutes ago. It can be seen that the correlation between the RR and the past RR decreases gradually over time, whereas the correlation to the past AL plateaus at 11 minutes and then declines further at around 20 minutes. Combining this with the desire to keep the model manageable, the choice was made to include 20 minutes of past RR and AL data as the TS inputs. This model will be referred to as the 'time series' (TS) model.

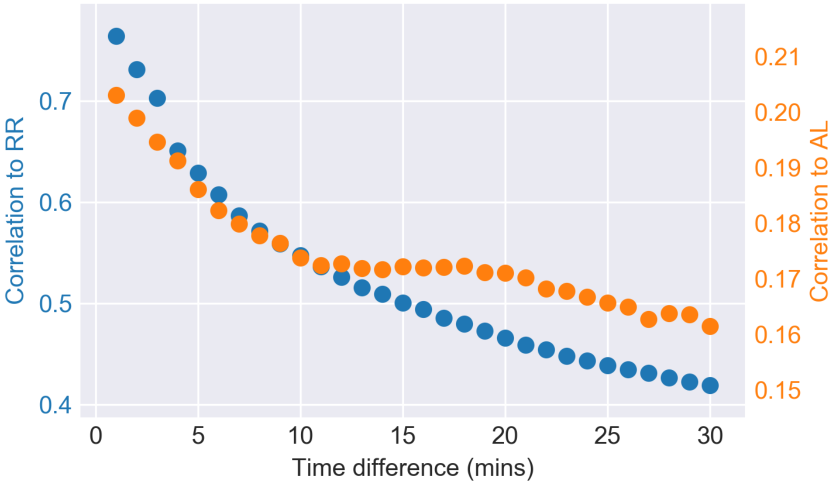


Figure 5.1: Correlation of current RR with past RR and AL (all subjects).

5.1.2.2 Time series and differences

The second model is an extension of the TS model. It is based on the idea that one of the key aspects of the RR development is the underlying trend or, in mathematical terms, what the first and second derivatives of the 'true' RR curve are. As the curve's mathematical formula is not known, but only samples at discrete times are available, this can be approximated by calculating first and second differences. However, as the RR itself is very variable, taking the first and second differences directly does not result in good indicators of the trend. Hence, the RR was smoothed before taking the differences. These were then included as inputs to the model. This model will be referred to as the 'TS + differences' model.

Figure 5.2 shows a sample of the RR of subject 60 along with the same signal smoothed by four degrees: 5, 10, 15, and 20 minutes. It can be seen that smoothing over 5 minutes results in a breathing signal that is quite variable. To a lesser extent, the same holds for smoothing over 10 minutes. The smoothing over 15 and 20 minutes produces less variable signals, but both have the disadvantage that the signal is slow to react to changes in the RR. However, as the main aim here is to represent the trend, it was decided to smooth over 15 minutes, as a compromise between trend representation and adaptation speed.

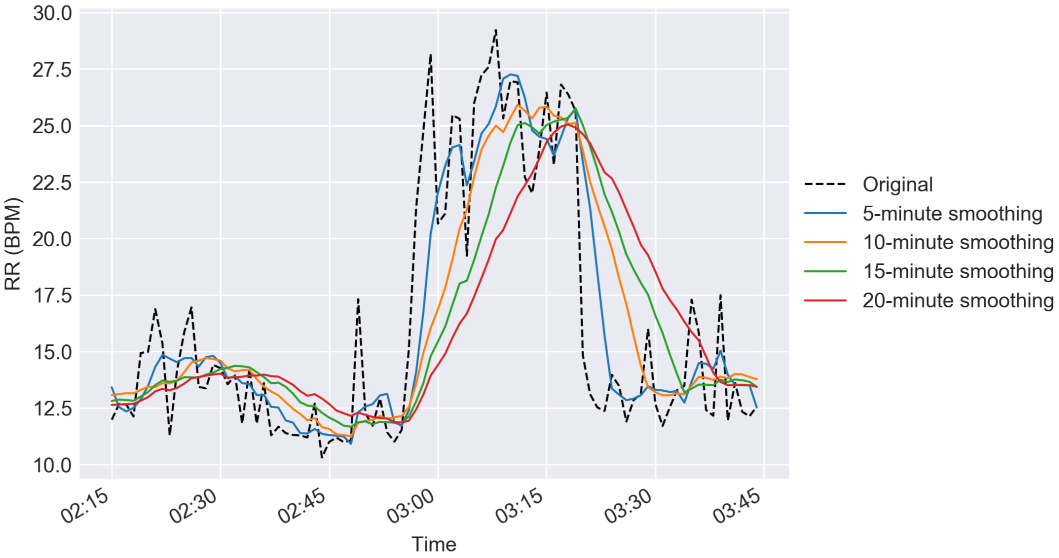


Figure 5.2: RR smoothed to different degrees (data sample from subject 60).

5.1.2.3 Time series and fixed features

The third model is also an extension of the TS model, this time with certain 'fixed' features, and will therefore be referred to as the 'TS + fixed' model. In contrast to the features discussed so far, which were all based on time series, the term fixed is used to refer to features that are constant over the input time window, and which consist of just one value. A distinction was made between two types of fixed features: subject characteristics and TS characteristics. The subject characteristics that were included are age, sex, BMI, and pulse. Each of these can affect the average RR of a person, as well as the way AL and RR interact, and should therefore be helpful to improve the predictions.

The TS characteristics include statistical characteristics about the time series inputs, inspired by Prasertsung and Horanont (2016). For both the RR and AL, the maximum, minimum, standard deviation, mean, and range in the input window were included as features. The largest increase and decrease in one minute were also added for both. Lastly, three features about the relationship between RR and AL in the past were added, namely their correlation over the input window, and the maximum and mean ratio of RR and AL. These features contain information about the extent to which the RR and AL ‘match’. The reasoning behind this is that as the RESpeck measures quiet breathing at rest, this would identify when the RR is reported as for periods of activity, and will be corrected in subsequent minutes. For example, when the AL was high but the RR was not, it can be expected that the RR will go up soon after in a delayed response.

5.1.2.4 Time series, differences, and fixed features

The fourth model uses all the aforementioned features to combine the different kinds of information that each of them was intended to capture. Although this seems like a logical thing to do, it comes with the risk of ‘overfeeding’ the model with too many features, making it difficult to train. The model will be referred to as the ‘TS + differences + fixed’ model.

5.1.3 Respiratory trend models

Preliminary results of the RR predictions suggested that the RR might be too variable to predict well. For the same reason, predicting the minute-by-minute RR directly may not be as useful as predicting the trend. These two considerations resulted in similar models as before, which instead estimated the RR trend as the outcome variable. This trend was represented by the smoothed RR, again over the past 15 minutes. The same four models were used, with the only difference that both the RR and AL inputs were changed to their smoothed counterparts, and so was the outcome variable. This task will be referred to as (respiratory) trend prediction, as opposed to the respiratory rate or RR prediction as described in the previous section.

5.2 Estimation and evaluation

5.2.1 Procedure

5.2.1.1 Evaluation metrics

In order to optimise the models, it is necessary to define metrics for good performance. The first is the Mean Squared Error (MSE), which is defined as the average of all squared predictions errors. Due to the squaring, the MSE punishes larger errors more than smaller ones. It was used both as the loss function for the optimisation algorithm (Section 5.2.2), and as the main criterion during hyperparameter tuning (Section 5.2.3).

Two other metrics were added to give a broader understanding of performance. The Mean Absolute Error (MAE) is similar to the MSE, except that instead of squaring the prediction errors to prevent positive and negative errors from cancelling out, their absolute value is taken. To enable easy comparison between the MSE and MAE, the root MSE (RMSE) is used from here on, which has the same scaling as the MAE.

The third evaluation metric was devised with the aim of capturing whether or not a model describes trends well. This is useful in addition to the RMSE and MAE, because a mean-based model may have a low RMSE/MAE, but after a few minutes its predictions will level off to a straight line, which is not very helpful. Therefore, the percentage of predictions that have the correct direction ('% correct direction') was calculated. A prediction was counted as 'correct' if it predicted an increase for a minute in which the RR had an increasing trend (i.e. the smoothed RR increased), or predicted a decrease/no change for a minute in which a decrease/no change in the trend was observed. The '% correct direction' is expected to be around 50 if predictions are made at random. Hence, a model does better at predicting the trend if its '% correct direction' is higher than 50. For the trend models, this statistic was calculated based on the smoothed RR directly. For the RR models, it was calculated on the smoothed RR shifted forward by eight minutes (approximately half of the 15 minuts used for smoothing) so that changes in the true RR are aligned with the smoothed signal, which naturally has a delay.

5.2.1.2 Training, validation, and testing

Having introduced the evaluation metrics, attention can now be given to the evaluation procedure. The concept of overfitting was explained shortly in Section 5.1.1 in the context of Ridge regression, as to how it controls overfitting through its loss function. However, another important technique to prevent overfitting lies not in the models themselves, but in the training and evaluation procedure. It is common practice to split up the dataset into three parts that are used for training, hyperparameter tuning, and final evaluation respectively, called the training, validation, and testing data.

The idea is that the candidate models are trained only on the training data for many different hyperparameter settings, after which the best hyperparameters for each model are chosen by comparing the performance on the validation data. Because these hyperparameters were chosen by looking at the validation data, the performance on the validation data is not generally an accurate estimator of the model's generalisation performance. The last step is therefore to use the selected models to make predictions on the testing data, which has not been seen before by the models and which has not been used to make any decision about the model settings. The performance on the test data is then used to compare the final performance of the models. This is the approach that was taken to estimate the LSTM models. Section 5.2.3 describes this in more detail.

Instead of splitting off one part of the data as validation data, it is also possible to do hyperparameter tuning using k-fold cross-validation (CV). In this case, the dataset is not split into three parts directly, but only into training and test portions. Subsequently, the training data is split into k parts, called folds, and the model is estimated k times, each time using $k-1$ folds as training data and the kth fold as validation data. The advantage of this approach is that no portion of the data is held out completely for validation, making more training data available. However, as each model and hyperparameter setting is estimated k times, it can be too time-consuming for models that take long to estimate. For this reason, it was not applied to the LSTM models, but was used to tune the Ridge regression.

5.2.2 Optimisation algorithm

Because there is not normally a closed-form solution to minimising the loss function of a neural network, training a neural network relies on the use of a numerical optimisation algorithm to find the loss-minimising parameters. There are many optimisers that can be used for this task, but gradient descent algorithms are the most common. The main reason for this is that second-order methods, such as Newton's method, require second derivatives, which are often infeasibly difficult to calculate for neural networks. The first derivatives, on the other hand, can be obtained relatively well using backpropagation, a method used to calculate the gradient of the loss function backwards through a neural network.

The idea behind gradient descent is to minimise an objective function by 'following' the surface of the function downhill until its lowest point by updating the parameters in the right direction, and the negation of the gradient tells us what the right direction is. Updates to the parameter vector θ are therefore made as:

$$\theta = \theta - \eta \cdot \nabla_\theta J(\theta) \tag{5.3}$$

where $J(\theta)$ is the loss function and η is the learning rate, a hyperparameter that is set by the user and determines the size of the parameter updates. Another hyperparameter that needs to be set is the batch size, which is the number of training examples for which the gradient is calculated before an update is done. The original (batch) gradient descent sets the batch size equal to the total number of observations in the data set, so that the gradient is always calculated with respect to all observations. The other extreme, called stochastic gradient descent (SGD), has a batch size of 1, updating the parameters by calculating the gradient with respect to one observation at a time. This is much faster than the standard gradient descent, but the objective function value varies a lot in the process and for that reason the algorithm may not be able to approximate the exact minimum well if the learning rate is kept constant. A compromise approach is mini-batch gradient descent, which uses a batch size anywhere in between these two extremes.

The main difficulty when using gradient descent for the optimisation of neural networks is that the loss functions are usually highly non-convex, which means that there are many local optima for the algorithm to get stuck in. Many algorithms also have trouble escaping saddle points, as the gradient surrounding these points is very close to zero (Ruder, 2016). It is therefore essential to pick an optimiser that can deal with these issues, and preferably exploit the specific characteristics of a dataset. An algorithm's ability to do this is dependent on the adaptations that are made compared to standard gradient descent. The two most important adaptations for neural networks are described below, following mainly the article by Ruder (2016).

The first important concept is that of momentum. The aim of momentum is to speed up learning when the gradient consistently points in the same direction and to slow it down when the gradient changes direction often. It is implemented by adding a fraction of the previous update vector to the current update vector, thereby increasing the size of the update for parameters that are moved in the same direction as in the previous iteration, and decreasing it for parameters that are moved in the opposite direction. The second adaptation is to keep a separate learning rate for all parameters, so that parameters that are updated infrequently receive larger updates. These learning rates are called adaptive, as they adapt to the update frequency of each parameter. They are particularly helpful if the data is sparse. This is not the case for the data in this project, but as algorithms with adaptive learning rates are also less sensitive to the start value of the learning rate, it is still expected to be a useful addition.

There are multiple algorithms implementing either momentum or adaptive learning rates in one way or another, but one algorithm combining them both has become even more popular: Adaptive Moment Estimation, or Adam short (Kingma and Ba, 2014). For the momentum, Adam keeps track of an 'exponentially decaying average of the past gradients', m_t. For the adaptive learning rates, it does the same, but with the squared past gradients, giving v_t (Ruder, 2016):

$$m_t = \beta_1 v_{t-1} + (1-\beta_1)\nabla_{\theta_t} J(\theta_t) \tag{5.4}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)\left(\nabla_{\theta_t} J(\theta_t)\right)^2 \tag{5.5}$$

where β_1 and β_2 are the decay rates. We can see that the magnitude of m_t increases if the gradient at time t points in the same direction as the gradient at time $t-1$, and that v_t increases more for parameters with a larger gradient. After a bias correction on m_t and v_t, the parameter updates for Adam using the corrected values, $\hat{m}_t$ and $\hat{v}_t$, become:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \varepsilon}\hat{m}_t \tag{5.6}$$

where ε is a very small term to prevent division by zero. From this update rule it is clear that indeed $\hat{m}_t$ speeds up learning when consecutive gradients point in the same direction, as intended by momentum, and $\hat{v}_t$ slows down learning for parameters that are updated often, as intended by the adaptive learning rates.

The developers of Adam show that their algorithm compares favourably with other common algorithms, and informal testing of different algorithms on the models estimated in this dissertation confirmed this. Therefore, the Adam algorithm was selected as the choice of optimiser. All neural networks were implemented using the python library *Keras* (Keras, 2018b) with a *TensorFlow* backend (TensorFlow, 2018). The learning rate was set to 0.002 as opposed to Keras's default value of 0.001 based on initial tests. The decay rates β_1 and β_2 were left at the values suggested by the developers, 0.9 and 0.999 (Kingma and Ba, 2014). The batch size was also experimented with informally, but did not seem to have a big impact, so it was set to 100, which is a fairly common value. RMSE was used as the loss function, as explained earlier. One final trick applied to improve convergence was to standardise all input features by subtracting the mean and dividing by the standard deviation, which makes the error surface easier to read for gradient descent algorithms.

5.2.3 Hyperparameter tuning

As a total of fourteen models had to be tuned (three potential baseline models and four LSTMs for both the respiratory rate and respiratory trend predictions), this section is split into four parts grouping together models that were tuned in a similar fashion. In each part, the tuning procedure is described in detail for one model and the final selection of hyperparameters is given for the others, with the supporting graphs provided in Appendix B.

5.2.3.1 n-period models

For both the n-period mean and slope the only hyperparameter to tune was n. Because these baselines are not prone to overfitting, they did not require a train/test split of the data. Instead, the RMSE was calculated directly on all data for different n. For the n-period mean, n was varied on a grid from 1 to 5. For the n-period slope, the grid ranged from 1 to 10. Figure 5.3 plots the RMSE of the n-period mean when used for RR prediction against the prediction window. It can be seen that except for the 1- and 2-period mean, there is barely a difference. n was therefore set to 3. A similar reasoning was followed for the n-period mean for trend prediction ($n = 1$), the n-period slope for RR prediction ($n = 10$), and the n-period slope for trend prediction ($n = 3$) based on the plots in Figure B.1.

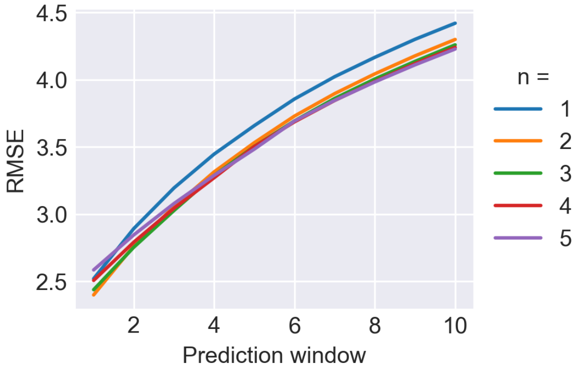


Figure 5.3: RMSE of n-period mean for RR prediction against prediction window for different n.

5.2.3.2 Ridge regression

For the Ridge regression baseline, it had to be decided both how many lags of the RR and AL should be included and what the penalty factor λ should be. For the former, a grid search was conducted over a 2D grid of values $\{0,1,2,5,10,20,30\}$, leading to 41 different lag structures/models, as zero lags for both the RR and AL is not a valid choice. For each combination, the optimal λ was determined through 5-fold CV on 80% of the data with the grid of options for λ set to $\{0.25,0.5,0.75,1.0\}$ after manually narrowing down the search to approximately this region. Each model was then estimated using the corresponding optimal λ and used to generate predictions for the held-out 20% test data. Finally, the RMSE on these predictions was used to determine the optimal lag structure.

Figure 5.4 summarises the information contained in the collection of RMSEs. It plots the RMSE averaged over the AL lags for all RR lag options, and vice versa. This shows that for the RR lags, there is a clear optimum at 5, regardless of the number of AL lags. For the AL lags, the minimum appears to be at 0, 1 or 10, but it is not as defined. A closer look at the RMSE values showed that a relatively simple model with 5 RR lags and no AL lags performed only marginally worse than the model with 5 RR lags and 10 AL lags that had the lowest RMSE (2.29 versus 2.22). On this basis, the lag structure $(5,0)$ was chosen for the sake of simplicity. The optimal λ for this model was 1. For the respiratory trend prediction, the lag structure was set to $(20,0)$ with optimal $\lambda = 0.5$ based on Figure B.2.

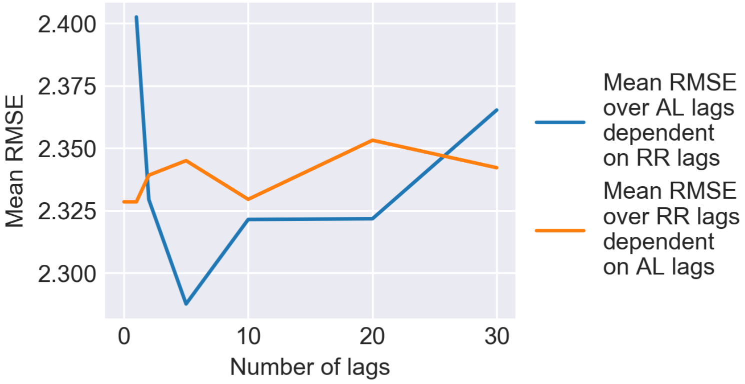


Figure 5.4: Mean test RMSE over AL/RR lags for each RR/AL lag option, Ridge regression for RR prediction.

5.2.3.3 LSTMs with TS inputs only

The LSTM models were optimised differently depending on the presence of non-TS features, because this required a different architecture of the neural network. This section describes the optimisation of the neural networks with only TS inputs, i.e. the TS and TS + differences models.

The optimisation was performed using a training, validation, and test set. These were created by splitting up the subjects rather than the recorded minutes, because if minutes recorded by the same subject are included in different sets, information on that subject is used for more than one task (‘data leakage’), defeating the purpose of splitting the dataset. The allocation of subjects to sets was at random in proportions 60:25:15. This resulted in 63 subjects being used for training, 27 for validation, and 16 for testing. This is fewer than the total number of subjects in the dataset, because not all subjects recorded night data.

The main decision to make was the architecture of the network. In order to obtain the desired RR and AL predictions, the networks always ended with a dense output layer of 2 units. Other than that, there was a free choice regarding the number of LSTM and other layers to use. For the model with only TS inputs, networks with one and two LSTM layers were tried out, with the number of units in each layer varied along the grid $\{3, 5, 10, 15, 20\}$. The 1-layer and 2-layer networks were trained for 150 and 300 epochs, respectively, to account for their complexity. In order to reduce computation time, training was stopped early if the decrease in validation loss was less than 0.002 for 15 epochs. Similarly, to prevent overfitting, a 20% dropout rate was set for the

LSTM layer(s), which means that during training, all units in the network have a 20% chance of being ignored for a forward/backward pass. This makes the network less prone to overfitting, because it requires it to learn more robust weights that also work if the network is changed slightly. The models were trained to predict one minute ahead using a batch size of 100.

Figure 5.5 summarises the validation RMSE as a function of the number of units in layer 1 and 2. The minimum is at $(10,3)$, closely followed by $(3,0)$ and $(5,0)$. Overall, an upward trend is visible with respect to the number of units in layer 2. This suggests that since the problem is relatively simple in terms of inputs, a large second layer overcomplicates things.

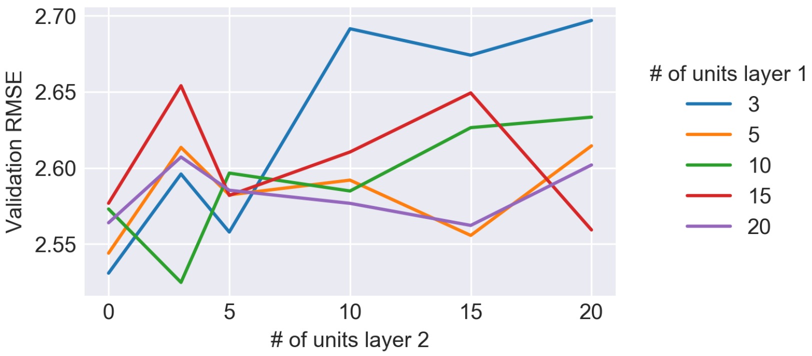


Figure 5.5: Validation RMSE as a function of the number of units in LSTM layer 1 and 2, TS model for RR prediction.

This upward trend is confirmed in Figure 5.6, which shows the mean RMSE when averaging over the number of units in layer 1 or 2. It can be seen on the right that both the validation and training RMSE are lowest on average for 0 units in layer 2, i.e. for a 1-layer network. The left graph instead shows a dip at 5 or 20 units in layer 1. Given that $(5,0)$ was the third best configuration in terms of RMSE, the final architecture selected was 1 LSTM layer with 5 units, or LSTM(5) - Dense(2) when considering the entire network. The optimal TS + differences models for RR prediction and trend prediction both were LSTM(3) - Dense(2). Finally, the architecture selected for the TS model for trend prediction was LSTM(3) - LSTM(3) - Dense(2). The corresponding graphs are included in Figure B.3 and B.4.

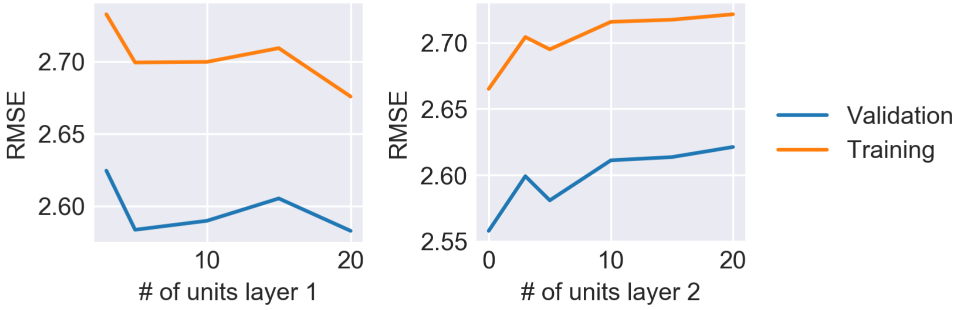


Figure 5.6: Mean RMSE over number of units in layer 2 (left) and layer 1 (right), TS model for RR prediction.

5.2.3.4 LSTMs with TS and fixed inputs

This section describes the optimisation of the neural networks with TS and fixed inputs, i.e. the TS + fixed and TS + differences + fixed models. The same train/validation/test sets were used as for the models without fixed inputs. Also, the dense output layer of 2 units is the same. There are differences when considering the rest of the network architecture. In particular, the models in this section needed to process both TS and fixed inputs. They were therefore modelled on the multi-input model described in the Keras functional API guide (Keras, 2018a), of which a schematic representation can be found in Figure 5.7. As there are two types of inputs, the models basically consist of two parts. First, the TS inputs are processed through one or more LSTM layers. The outputs of this part of the network are then merged with the fixed features, after which the second part of the network processes them all together to obtain the final outputs.

For each of the models in this section, there is a model from the previous section that essentially corresponds to the first part of the network. The model tuning was therefore sped up by fixing the first part of the network to the optimal architectures found for the corresponding model without fixed features. For example, the first part of the TS + fixed model for RR prediction was set to LSTM(5), because that was the optimal architecture found for the TS model for RR prediction. Figure 5.7 shows that the first part of the multi-input network returns not only the processed TS features, but also an auxiliary output. The documentation states that this helps to train the LSTM smoothly (Keras, 2018a). The same approach was therefore adopted, weighing the main output loss by a factor 1 and the auxiliary output loss by a factor 0.2.

This left only the second part of the network to be tuned. This was done by adding one

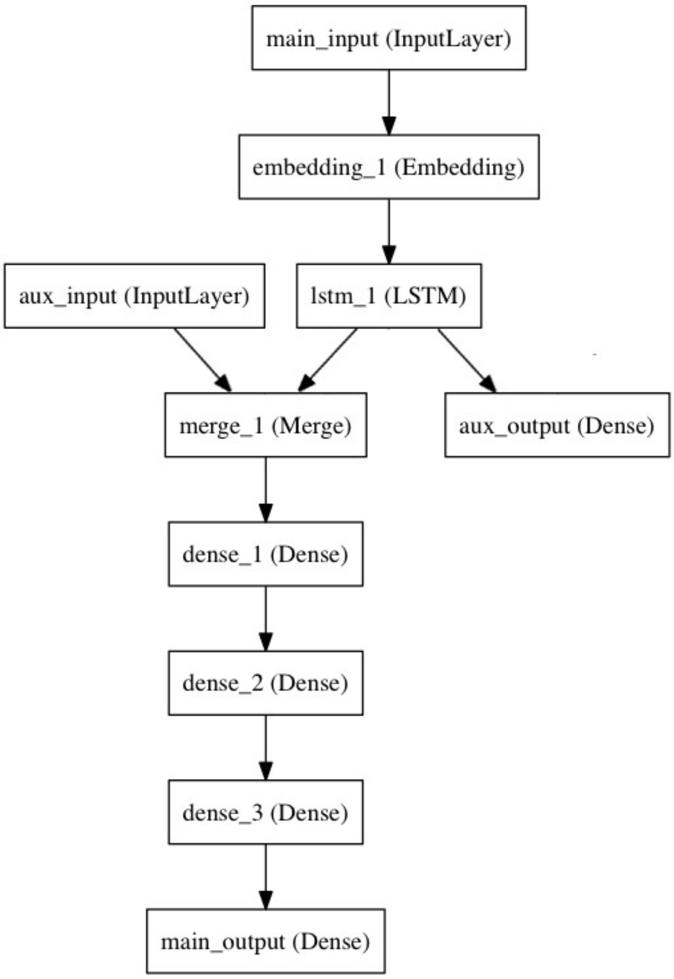


Figure 5.7: Example of a multi-input model. Taken from Keras (2018a).

or two more dense layers after merging the (processed) inputs. The number of units for these layers was varied on the grid $\{3,5,10,20\}$, where it was required that the number of units in the first layer exceeded the number of units in the second layer so as to force the network to gradually reduce the dimension of the data to the required 2D output vector. Like the models with only TS inputs, the networks with a 1- and 2-layer second part were trained for 150 and 300 epochs respectively and training was stopped early if the decrease in validation loss was less than 0.002 for 15 epochs. The same 20% dropout rate was also applied to all layers.

Figure 5.8 shows the validation RMSE for the TS + fixed model on RR prediction dependent on the number of units in layer 1 and 2, where layer 1 and 2 refer to the dense layers in the network's second part. It can be seen that the optimum for layer 2 is clearly at 3 units, possibly combined with 20 units in layer 1. Figure 5.9 again plots the mean RMSE when averaging over the number of units in layer 1 or 2. The plot on the left shows that the number of units in layer 1 should be either 3 or 20, so given that the lowest validation RMSE was achieved for $(20, 3)$, the architecture LSTM(5) - Dense(20) - Dense(3) - Dense(2) was selected as optimal.

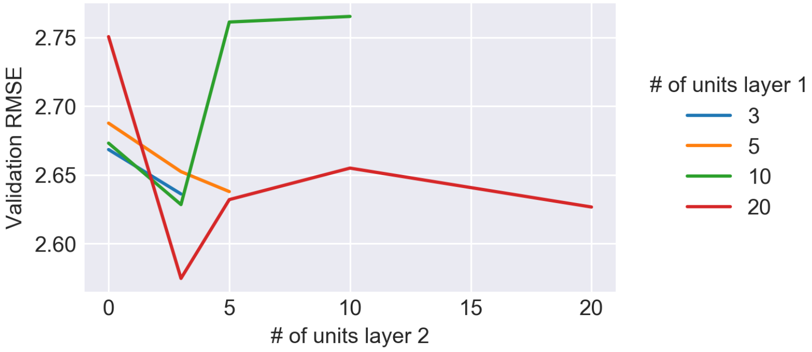


Figure 5.8: Validation RMSE as a function of the number of units in LSTM layer 1 and 2, TS + fixed model for RR prediction.

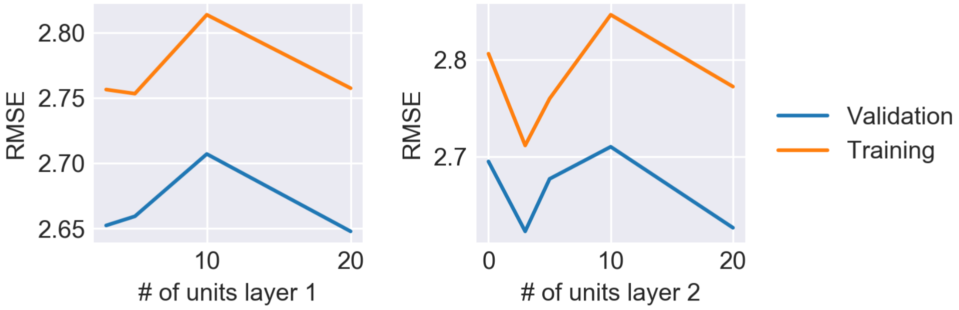


Figure 5.9: Mean RMSE over number of units in layer 2 (left) and layer 1 (right), TS + fixed model for RR prediction.

The final architecture for the TS + fixed model for trend prediction was LSTM(3) - LSTM(3) - Dense(10) - Dense(3) - Dense(2). For the TS + fixed + differences model for RR and trend prediction, they were LSTM(3) - Dense(20) - Dense(20) - Dense(2) and LSTM(3) - Dense(5) - Dense(2) respectively. The corresponding graphs are included in Figure B.5 and B.6.

5.3 Results

The estimated models were used to recursively predict the RR from one to ten minutes ahead, meaning that each time, the most distant minute of true data was removed from the inputs, and the previous minute's prediction was added. This section presents the results of these predictions first for the respiratory rate and then for the respiratory trend models.

5.3.1 Predicting respiratory rate

Table 5.1 shows for all models, sorted by RMSE, the average test MAE and RMSE over the 1- to 10-minute prediction windows, and the percentage of predictions where the direction was correct. The lowest RMSE is achieved by the TS model, followed by the TS + fixed model. The lowest MAE is achieved by the 3-minute mean, followed by the TS + fixed and TS models. For the '% correct direction', the TS + fixed and TS models are beaten only by the Ridge regression.

Table 5.1: Average test performance of respiratory rate models over all prediction windows, sorted by RMSE.

	MAE (BPM)	RMSE (BPM)	% correct direction
TS	2.244	3.167	55.0
TS + fixed	2.218	3.199	55.2
TS + differences	2.292	3.246	49.7
3-minute mean	2.207	3.304	49.0
TS + differences + fixed	2.586	3.446	53.5
Ridge regression	2.729	3.559	57.2
10-minute slope	3.499	4.991	46.6

The TS + differences + fixed model does not perform well, outperforming only the Ridge regression and 10-minute slope. This might be due to irrelevance of the 'differences' features, because the TS + differences model also does not do too well. Although a neural network has the option to set the weights of these features to zero if it finds they are irrelevant, determining the relevance can be difficult. One reason why these features may be irrelevant is that the differences were taken over the smoothed RR, but the actual RR is so variable that knowing the derivative of its trend may not be very helpful. Given their good performance, the 3-minute mean (baseline), TS, and TS + fixed models are discussed in more detail.

Figure 5.10 plots the MAE, RMSE, and '% correct direction' per model against the prediction window length. It is clear and unsurprising that predicting further ahead is more difficult. It can also be seen that although the average MAE is technically lowest for the 3-minute mean, the curves are practically the same when plotted against the prediction window. The RR can be predicted quite well; the average error is about 1.5 BPM for predictions one minute ahead and increases to 2.7 BPM for ten minutes ahead.

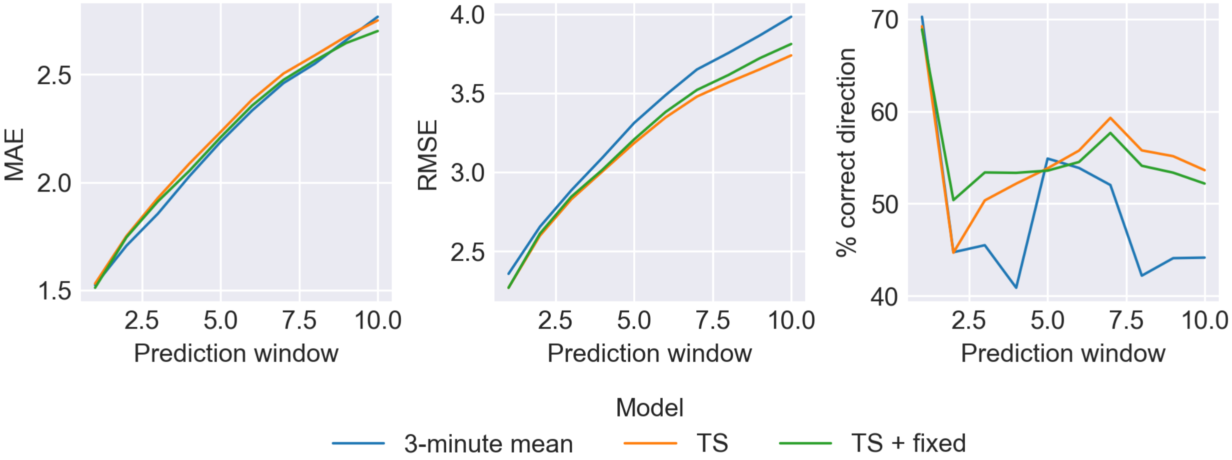


Figure 5.10: Average MAE (left), RMSE (middle), and % correct direction (right) of respiratory rate models as a function of the prediction window length.

For the RMSE, the differences are more substantial, and they grow with the prediction window to about 0.25 in favour of TS and TS + fixed when predicting ten minutes ahead. This is not surprising given the fact that the 3-minute mean model will start returning the same value when it has to predict more than a few minutes ahead. This is demonstrated in Figure 5.13a, which plots the true RR of subject 29 against a sample of 10-minute predictions of the three models. It can be seen that all the 3-minute mean predictions level off after a short while. The plot also shows that the 3-minute mean is not very good at handling outliers, which explains its higher RMSE overall. The prediction starting between 1:30 a.m. and 1:45 a.m. shows how the observed RR drops very steeply from over 30 BPM to only 20 BPM, an outlier, which is detected to some extent by both the TS and TS + fixed models, but not by the 3-minute mean.

The '% correct direction' shows interesting patterns for all three models. They have a good performance for one minute ahead, then a sharp drop to below 50% even for the TS model and 3-minute mean. This is followed by an increase as the prediction window increases, although it also drops again for the 3-minute mean. A good explanation for this pattern could not be found.

Comparing the TS model to the TS + fixed model, there are minor differences in the MAE, where TS + fixed fares better, and the RMSE, where TS does better. This sounds logical when one considers the features of each model. The fixed features include information about the subject that relates to their average RR, such as BMI and resting pulse. This allows the model to adapt its predictions to a person's 'normal' RR, which should make them more accurate, resulting in a lower MAE. At the same time, the TS model has to rely purely on the time series information, making it better at detecting outliers, which may be signalled very subtly in the minutes before, resulting in a lower RMSE.

Having evaluated the models' overall performances, the analysis proceeds to investigate the predictions per subject in more detail. Figure 5.11 shows the average RMSE per model for all the test subjects. It can be seen that although the average performance is different per model, the relative performance on individual subjects is very similar, i.e. certain subjects are among the most poorly predicted regardless of the model (e.g. 73, 88), whereas others are always among the best (e.g. 26, 53). To determine what is driving this, the standard deviation of the RR per subject was plotted against their average RMSE of the three main models in Figure 5.12. The plot shows a clear positive relation between the two variables, indicating that an RR with a higher variance is harder to predict than a steady one.

To visualise the difference between a high- and low-variance respiratory signal, the night for subject 88 (RR standard deviation 4.97) was plotted in Figure 5.13b. This can be compared to the plot in Figure 5.13a of subject 29 (RR standard deviation 2.48), whose average RMSE is consistently in the lower half for all models. For subject 29, there is only one real elevation of the RR in the whole night. Subject 88, on the other hand, shows a very different breathing profile with multiple clearly separated episodes of quick breathing. The predictions within the slow- and fast-breathing periods look reasonable, and drops in the RR also seem to be predicted quite well, although the timing is not always accurate. More problematic are the predictions when a switch to fast breathing is about to occur, for example shortly before 2 a.m. and around 3:30 a.m. None of the models predicted these changes, which is not surprising given their extreme suddenness, but disappointing nonetheless with the potential application of monitoring patients' health status in mind. It is likely that these minutes are one of the

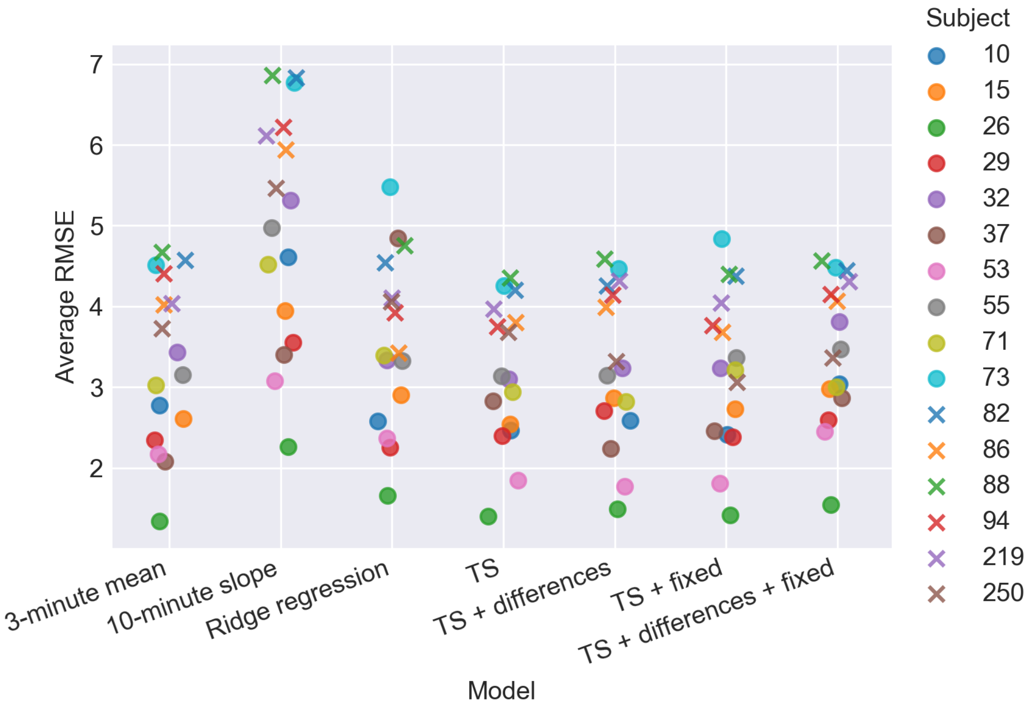


Figure 5.11: Prediction errors by model and test subject (direct RR).

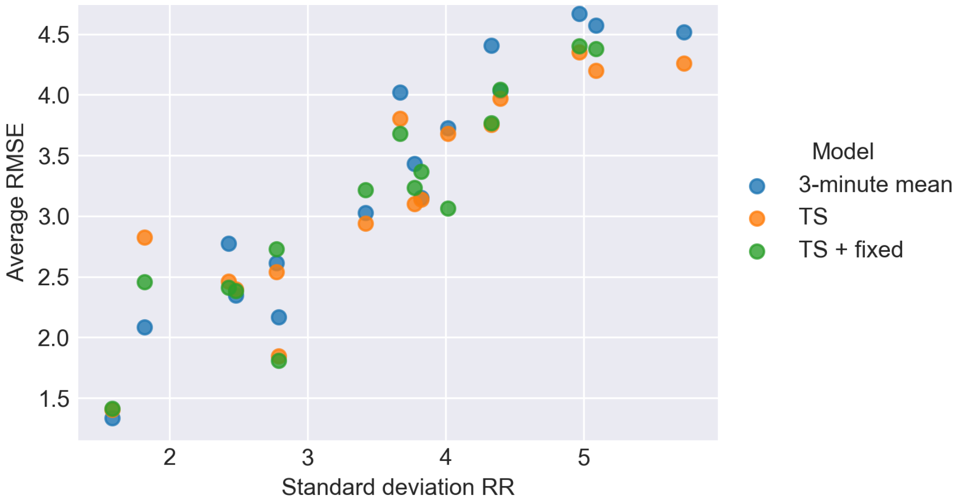
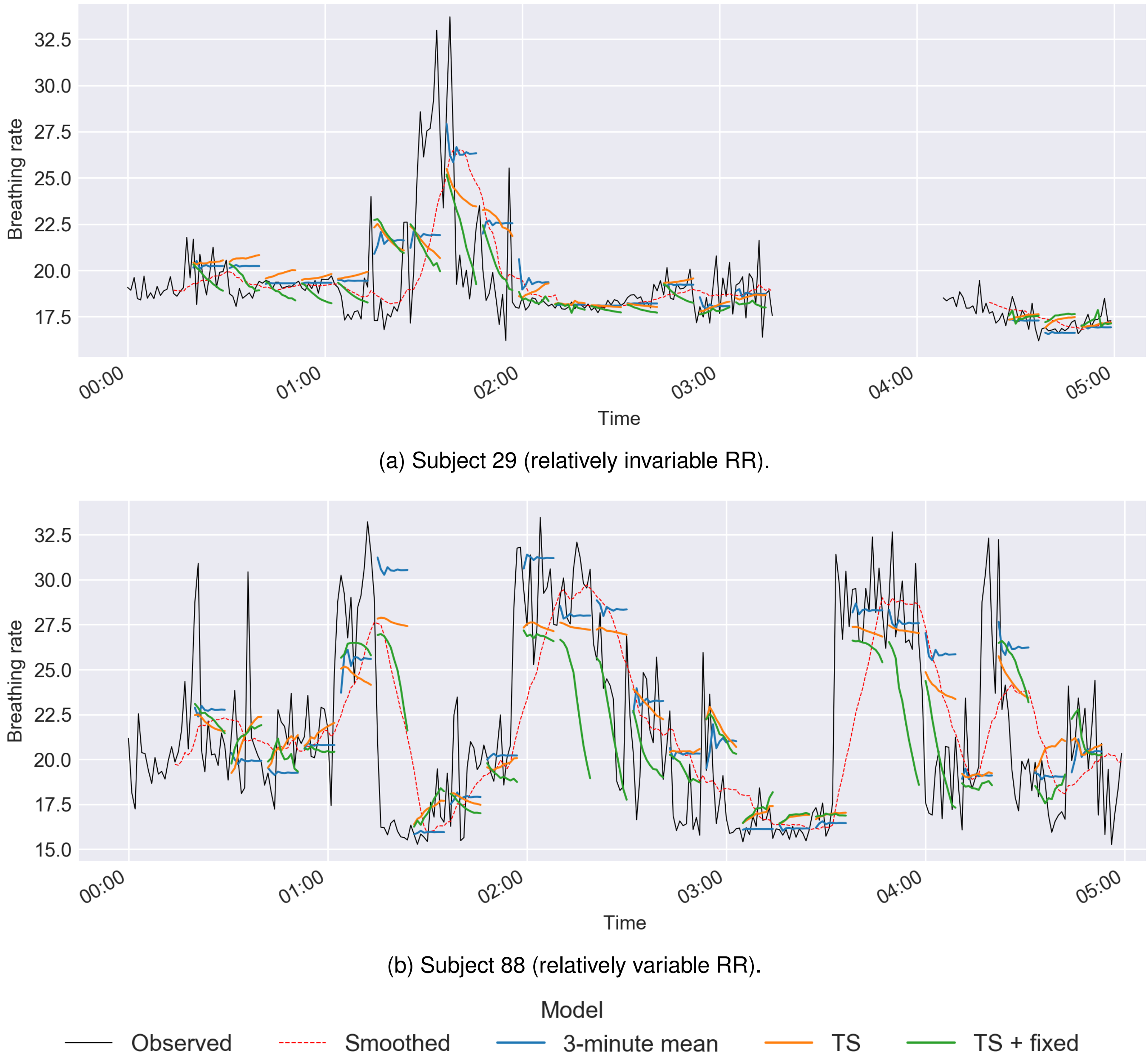


Figure 5.12: Average RMSE on RR prediction against standard deviation of the RR per test subject.



(a) Subject 29 (relatively invariable RR).

(b) Subject 88 (relatively variable RR).

Figure 5.13: Samples of 10-minute RR predictions against true RR.

main reasons for this subject's high average RMSE, because the undetected increases are so large; from under 20 to over 30 BPM.

Examining the particular respiration pattern in the plot, a thought that comes to mind is that they may be related to the subject's sleep phases. Two clear cycles consisting of a fast and slow breathing period can be seen, both around 90 minutes long (2:00 - 3:30 a.m. and 3:30 - 5:00 a.m.), which is also the approximate length of a sleep cycle. Moreover, it is known that breathing is different during different sleep phases (Riley et al., 1985). It would be interesting to investigate this hypothesis further, but more specific information on sleeping patterns is needed to do so. If it is indeed the case that the RR is so clearly dependent on the sleep cycle, the prediction model could benefit greatly from incorporating information about the sleep phase.

5.3.2 Predicting respiratory trend

The results for the respiratory trend models are presented in Table 5.2, sorted by '% correct direction'. The picture here is a bit more complicated than for the RR predictions. First of all, it should be noted that the smoothed RR is a lot easier to predict than the direct RR: the best RMSE decreased from 3.167 to 1.324. This is an improvement that is substantially larger than just the reduction in the target's standard deviation from 4.59 to 3.90. The TS + fixed model is on top based on the '% correct direction', but does poorly on both the MAE and the RMSE. The two simplest baseline models both perform well, in particular the 3-minute slope, because the 1-minute mean is not good at predicting the correct direction. The model that is able to compete with the 3-minute slope best is the TS + differences + fixed model. These two will therefore be examined in more detail.

Figure 5.14 again plots the MAE, RMSE, and '% correct direction' for the two selected models against the prediction window. It can be seen that the 3-minute slope beats the TS + differences + fixed model until about 8 minutes ahead for the MAE and 7 minutes ahead for the RMSE and '% correct direction'. An interesting fact is that the MAE for these models increases more with the prediction window than when the RR was predicted directly. Here, the MAE triples when comparing 1-minute ahead predictions to 10 minutes ahead. For the RR predictions, it did not even double.

Figure 5.15 shows the average RMSE per model and subject. The main question of in-

Table 5.2: Average test performance of respiratory trend models over all prediction windows, sorted by % correct direction.

	MAE (BPM)	RMSE (BPM)	% correct direction
TS + fixed	1.905	2.382	60.1
3-minute slope	0.955	1.357	58.2
TS + differences + fixed	1.056	1.393	56.6
TS + differences	1.178	1.442	55.6
TS	1.231	1.560	55.5
1-minute mean	0.946	1.324	49.5
Ridge regression	1.855	2.314	49.5

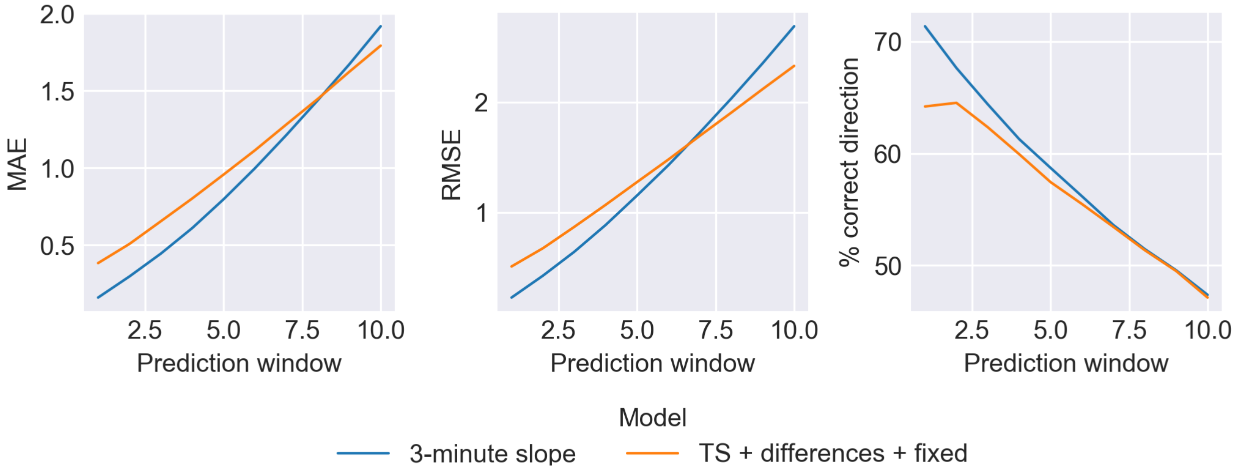


Figure 5.14: Average MAE (left), RMSE (middle), and % correct direction (right) of respiratory trend models as a function of the prediction window length.

terest here is whether the variance of the RMSE per subject decreased, because smoothing the RR means that the standard deviation of the subjects' RR is reduced, and this was the most important factor causing high RMSEs for some subjects in the previous section. The plot suggests that this question can be answered affirmatively, because the spread of the average RMSE looks smaller, and comparing the standard deviation of the RMSE per subject for the trend and RR models confirms this. Nevertheless, the same relative ordering of the subjects can still be observed, and plotting the standard deviation of the RR against the average RMSE per subject shows that the variance of the RR is still the driving factor behind differences in performance (Figure 5.16).

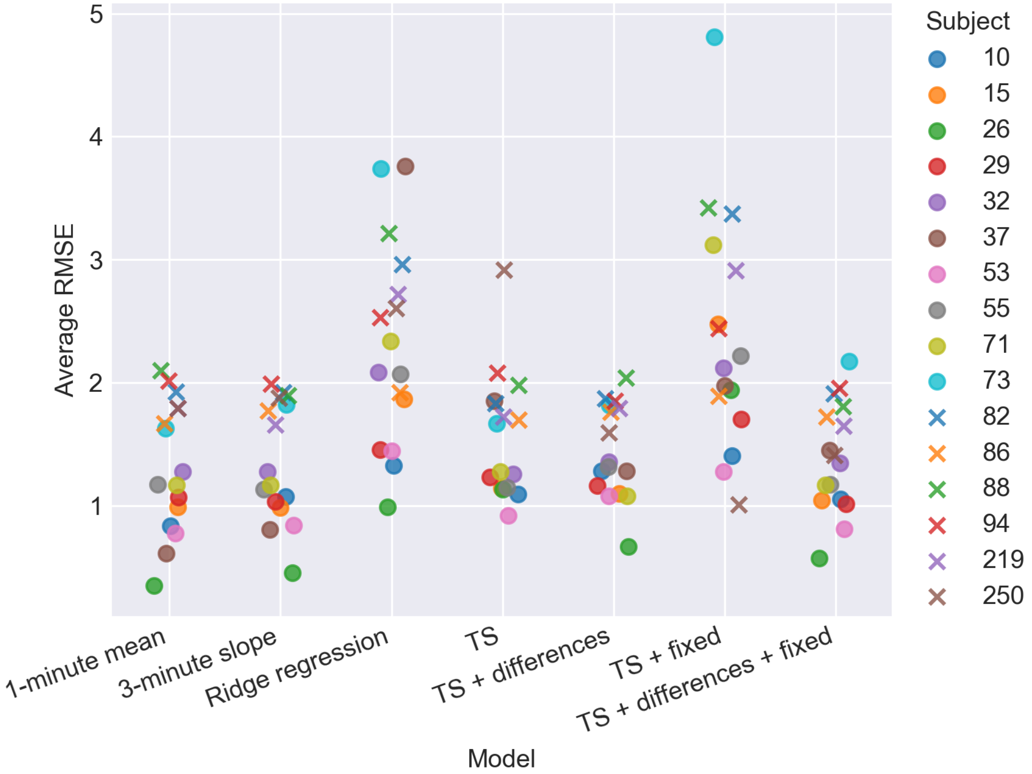


Figure 5.15: Prediction errors by model and test subject (smoothed RR).

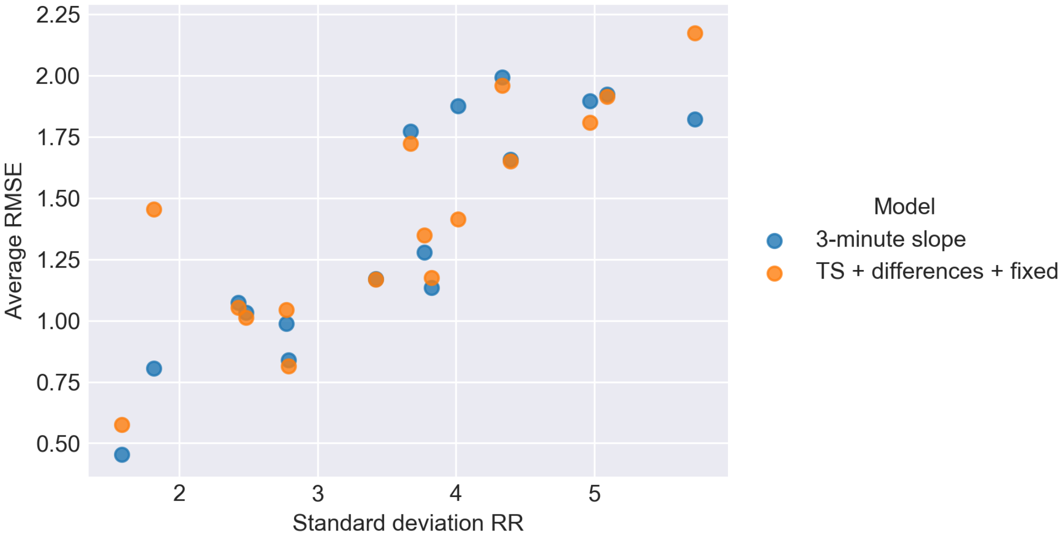
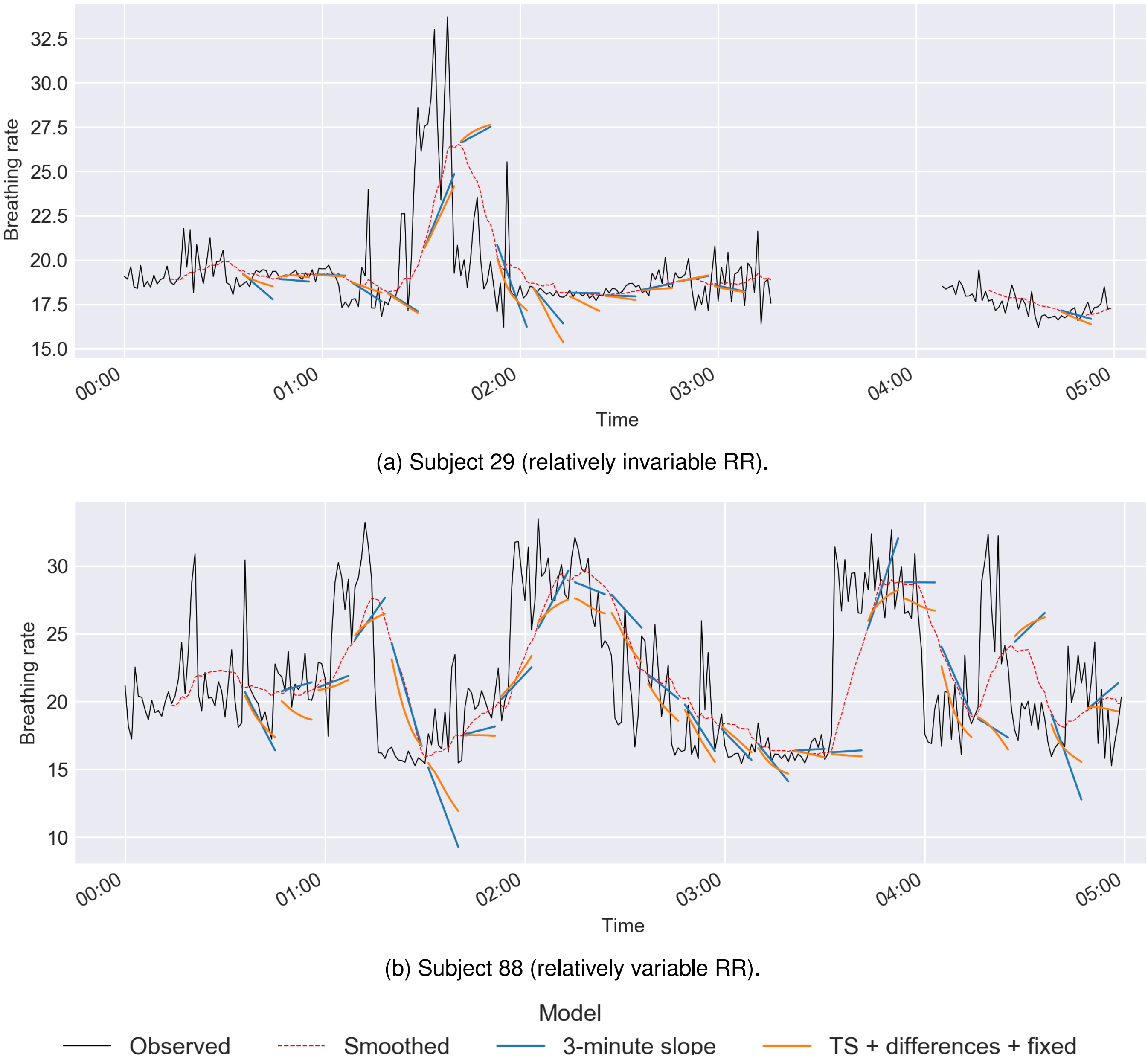


Figure 5.16: Average RMSE on trend prediction against standard deviation of the RR per test subject.

To compare the actual predictions visually, a sample of 10-minute trend predictions was plotted for the same two subjects as before: subject 29 with a low-variance RR

(Figure 5.17a) and subject 88 with a high-variance RR (Figure 5.17b). These plots show that the trend predictions for the baseline and TS + differences + fixed models are similar and both quite good. There are, however, some instances with big errors, occurring mostly at peaks and valleys. The TS + differences + fixed model seems to do better at these points, judging for example from the predictions for subject 88 starting around 03:40 a.m. (peak) and 04:40 a.m. (valley). While the 3-minute slope continues straight on in both cases, the TS + differences + fixed predictions show the beginning of a peak and a valley, respectively.



(a) Subject 29 (relatively invariable RR).

(b) Subject 88 (relatively variable RR).

Figure 5.17: Samples of 10-minute trend predictions against true RR.

Chapter 6

Effect of air pollution on respiration

A factor that can be expected to affect the respiratory rate but is missing in the APCAPS data, is exposure to air pollution. This chapter explores the relationship between air pollution and respiration based on a different dataset, and investigates how information on air pollution could be used to improve the RR predictions for patients diagnosed with asthma and COPD.

6.1 Relation between air pollution and respiratory rate

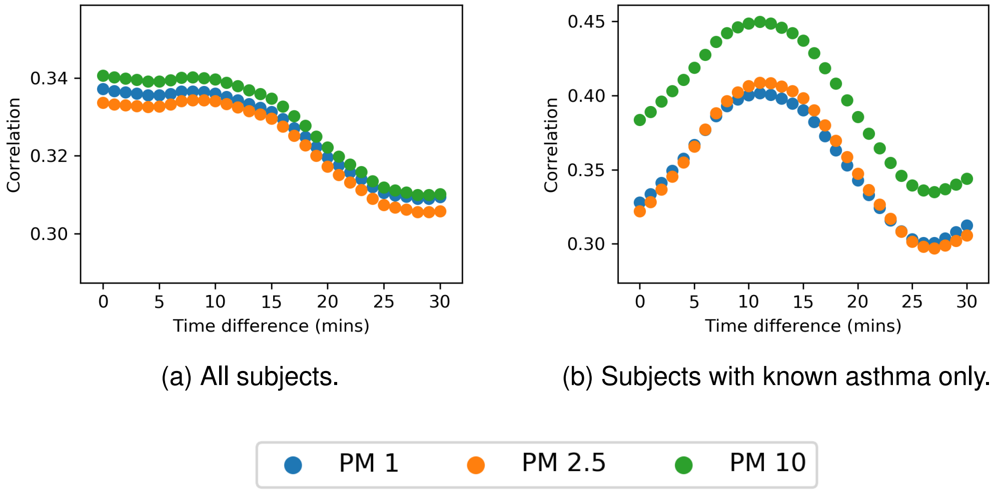
A study which includes data on both respiration and air pollution is Tackling Second-hand Smoke (TackSHS), a European research project in which the effects of second-hand smoking are investigated in Dublin (Ireland), Madrid (Spain), and Liberec (Czech Republic). It uses the same RESpeck sensor as in the APCAPS study to collect data about respiration and activity, but additionally its participants, who all have a respiratory condition, wear an *AirSpeck* belt, which measures the personal exposure to air pollution expressed in concentrations of particulate matter (PM) values (Arvind et al., 2016). Three types of PM-values are distinguished based on the diameter of airborne particles: PM1 refers to the concentration ($\mu g/m^3$) of particles smaller than 1 µm, PM2.5 and PM10 to the concentration of particles smaller than 2.5 µm and 10 µm, respectively. At the time of writing, data from eight of the first TackSHS subjects was available for analysis (Table 6.1). This data was used for an initial exploration of the question as to how air pollution affects respiration and whether this could be used to improve the RR predictions for patients with asthma or COPD.

Table 6.1: Known details on TackSHS subjects.

Subject ID	Country	Gender	Age	Diagnosis	Smoking status
CA01	Czech Republic	M	46	Asthma	Non-smoker
CC01	Czech Republic	M	71	COPD	Current smoker
SA03	Spain	M	25	Asthma	Non-smoker
SA04	Spain	F	55	Asthma	Current smoker
SA05	Spain	F	50	Asthma	Non-smoker
SA06	Spain	-	-	Asthma	-
SA07	Spain	-	-	Asthma	-
SC01	Spain	-	-	COPD	-

First of all, the direct correlation between the RR and the PM-values was analysed. Figure 6.1 shows the correlation between the current smoothed (15 minutes) RR and PM-exposure values for up to 30 minutes ago, aggregated over all subjects (6.1a) and only for subjects with known asthma (6.1b). The separation between COPD and asthma is important, because there are substantial differences between the symptoms of these diseases, so it is interesting to see the relationship for these patients separately. As there were only two COPD patients available, a separate plot was only created for the asthma patients. A disadvantage of aggregating the data could be that subjects who recorded more data have a larger influence on the computed correlations. However, this was preferred over calculating the correlations per subject and averaging them in this case, because it meant that subjects with few relevant minutes did not have to be excluded for unreliability of the correlation estimate. Given the limited number of subjects, this was judged to be more desirable than weighting all participants exactly equally. Additionally, a plot per subject is shown in Figure 6.2.

From the aggregated plot, it can be seen that for all types of particle pollution, there is an effect on the RR up until about 15 minutes after exposure. After that, the correlation drops slightly. For the asthma patients, there is an even clearer, positive correlation with the RR, which peaks between 10 and 11 minutes after the exposure. This pattern can be explained, as the latency is due to the effect taking time to manifest itself in the form of changes in the RR. The per-subject plots exhibit a high degree of individuality in the exposure-response relationships. Subject SA04 appears to be a very ‘average’ asthma patient in the sense that their response is similar as the average case. Subjects



(a) All subjects.

(b) Subjects with known asthma only.

Figure 6.1: Correlation of current RR with past PM-values (both smoothed).

SA03 and SA06 react almost instantly to exposure, wheras SA03 seems to have a longer time constant compared to the other asthma patients. For subject SC01, the peak response comes much later, around 25 minutes after the exposure. The plots for subjects CA01, SA05 and SA07 are almost counterintuitive, because there is no correlation or even a negative one to PM1 and PM2.5. A possible explanation is that their medication mitigates the effect of air pollution. Finally, the plots for subject CC01 and SC01, the only known COPD patients, show a lower correlation in general than in the case of asthma subjects, which matches the disease's symptoms.

To determine whether the correlations between RR and air pollution are direct relations or driven by confounders, the minutely RR was regressed on the PM-values from 11 minutes ago, based on the fact that the correlation for asthma patients shows a clear peak at a lag of 11 minutes, and a set of control variables. Firstly, binary indicators for the subject were included to capture variations in the average RR between subjects caused by factors such as fitness and age. Variation caused by activity levels were accounted for by including the activity type classification instead of the activity level directly, because the activity types are easier to interpret. The regression was based on the minutes during which the subjects were sitting/standing, lying down, or walking, because the RR during other activity types, such as movement, is unreliable. Lying down and walking were included as binary indicators, so that their coefficients should be interpreted as the change relative to sitting/standing. Finally, the recent history of respiration was accounted for by including lags of the RR. Only the odd lags from 1 to 9 were included, because otherwise the high autocorrelation of the RR caused multicollinearity-related issues with the estimation of standard errors.

Table 6.2 contains the results of the regression. The coefficients of all control variables are highly significant, except for the activity type 'walking'. The RR lags have positive coefficients that decrease with the lag. Minutes during which the dominant activity was walking are expected to have a higher RR compared to sitting/standing, and minutes during which lying down dominated are expected to have a lower one.

Considering the coefficients of interest, namely those of the PM-exposure values, PM1 and PM2.5 are highly significant, but PM10 is not. The coefficient for PM1 is negative, which seems counterintuitive at first sight, because an increase in PM1 ('pollution') is normally expected to lead to higher RRs. However, it is not strange when considering the definition of these variables: PM1 is a subset of PM2.5. This means that when interpreting the coefficient as the result of a *ceteris paribus*, one-unit increase to the PM1-value, we are effectively seeing the effect of increasing the *share* of PM1 in the amount of PM2.5, because the *ceteris paribus* condition includes the assumption that the total amount of PM2.5 stays the same. Hence, the coefficient tells us that for a certain amount of PM2.5, the RR is lower in minutes where the share of PM1 is higher. Interpreted more freely, this suggests that particles with a diameter between 1 and 2.5 µm increase the RR more than particles with a diameter less than 1 µm. The insignificant coefficient of PM10, seemingly contradictory to the correlations described earlier, can also be explained by this mechanism; an increase in PM2.5 leads to a higher RR, but an increase in PM10 with the same level of PM2.5 (i.e. an increase caused by particles between 2.5 and 10 µm) does not.

6.2 Using air pollution to improve respiratory rate predictions

Overall, it can be concluded that air pollution, in particular particles with a diameter between 1 and 2.5 µm, is positively related to the RR, even after controlling for other influences. This leads to the question whether the RR predictions could be improved using information on pollution levels. The eight subjects available from TackSHS are not enough to estimate a completely new model, so exploring this hypothesis requires a bit more creativity. The approach taken was to use one of the models estimated on the APCAPS participants, who are assumed to be representative of the general population,

Table 6.2: Regression of respiratory rate on air pollution and controls.

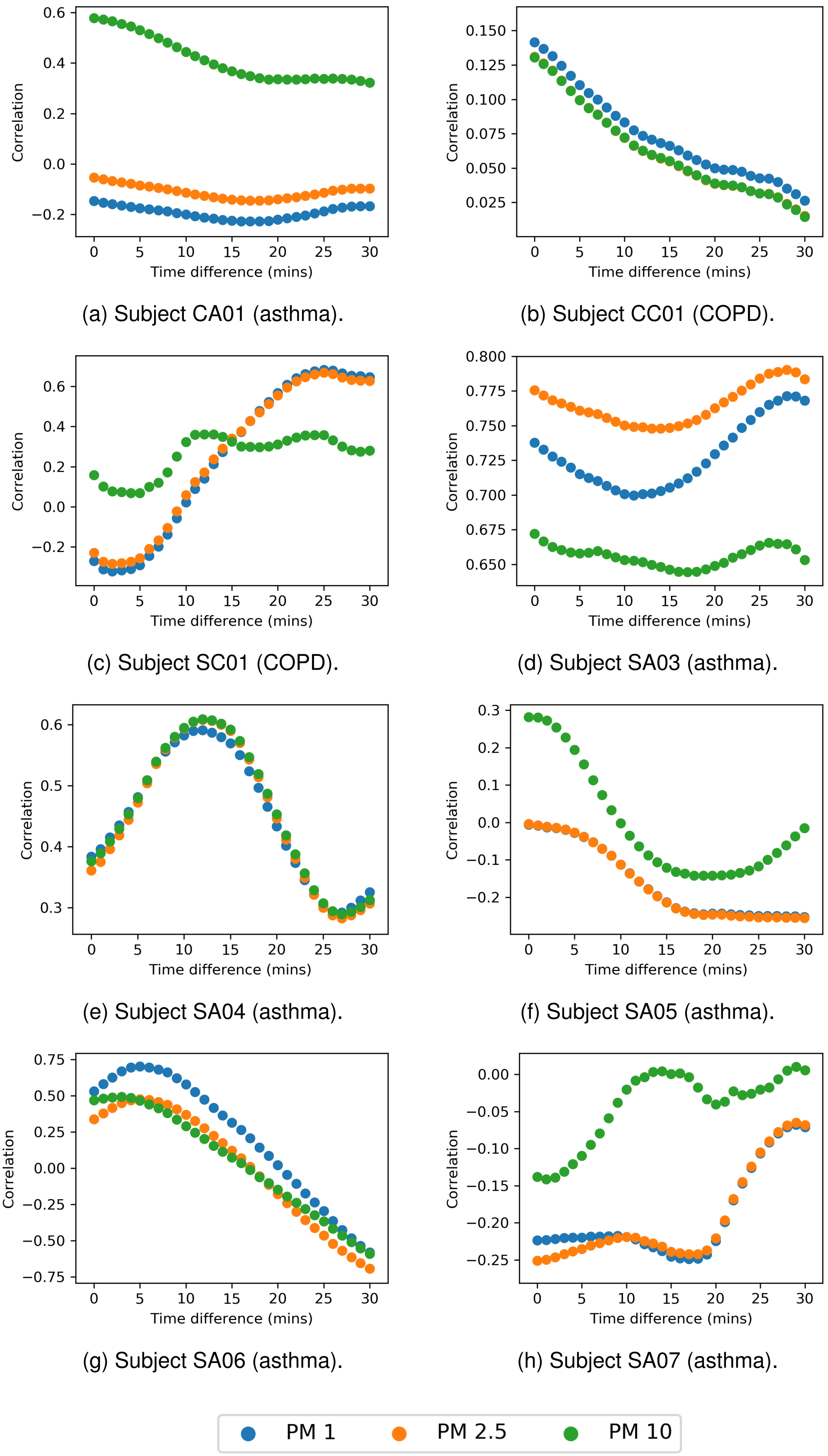
	Coeff	Std. error	P-value
Lying down	−0.171	0.09	0.048
Walking	0.803	0.62	0.196
RR lag 1	0.329	0.02	0.000
RR lag 3	0.221	0.02	0.000
RR lag 5	0.120	0.02	0.000
RR lag 7	0.099	0.02	0.000
RR lag 9	0.089	0.01	0.000
PM1 lag 11	−0.044	0.02	0.006
PM2.5 lag 11	0.037	0.01	0.007
PM10 lag 11	0.001	0.00	0.552
R^2: 0.68			
Observations: 3819			

to generate RR predictions for the TackSHS subjects, who have respiratory problems. The prediction errors could then be related to the air pollution levels. If air pollution adds value to the RR predictions over the information already contained in the model, a relation between the prediction errors and the PM-values should be visible. More specifically, the model is expected to underestimate the RR for higher pollution levels. The model chosen was the TS neural network for RR prediction, as it was one of the two best-performing models for this task.

Figure 6.3 plots the prediction errors (predicted minus true) against the PM-values after removing two outliers with PM-values over 150, first for all subjects together (6.3a) and then coloured by subject (6.3b). When considering all subjects together, a negative relationship is observed for all types of PM-values, which confirms the hypothesis that air pollution provides additional information about the RR on top of what is already learned by the model, because a negative prediction error means that the predicted RR was lower than the true RR. When considering the subjects individually, the picture is more complex, just like it was for the correlation plots (Figure 6.2). It can be seen that for subjects SA03, SA04, SA06, CA01, and CC01, the negative relation also holds individually. For subjects SA05, SA07, and SC01, it does not. This can be explained when scrutinising the correlations for these three subjects at a lag of 11 minutes (Figure 6.2f, 6.2h, and 6.2c): they are practically zero or negative.

Caution should be exercised when interpreting these results, because the TackSHS data is quite different from the data used to train the APCAPS model. This lies partly in the type of subjects, i.e. differences in ethnicity and health status (pulmonary disease versus assumed healthy), and partly in the type of data used, i.e. estimation on night data, but applied to combined day and night data. The model is therefore less expected to work well with the TackSHS subjects.

Nevertheless, two conclusions can be drawn from the analysis carried out in this chapter. Firstly, air pollution has an impact on the RR of people with pulmonary disease. This is worth investigating further both in the context of RR prediction and health effects. Secondly, the effect of air pollution on the RR is highly variable from person to person. The data that will be collected under the TackSHS project in the future should help shed light on the exposure-response relationships in more detail.



(a) Subject CA01 (asthma).

(b) Subject CC01 (COPD).

(c) Subject SC01 (COPD).

(d) Subject SA03 (asthma).

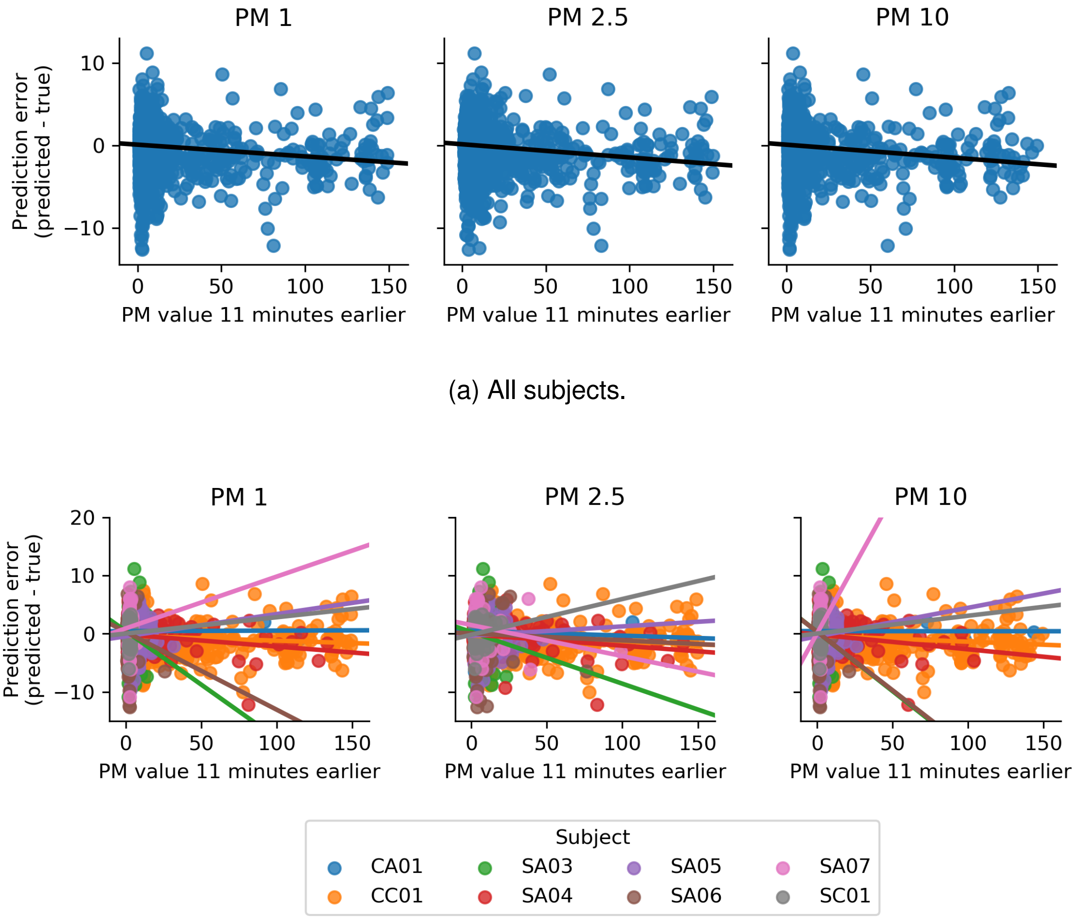
(e) Subject SA04 (asthma).

(f) Subject SA05 (asthma).

(g) Subject SA06 (asthma).

(h) Subject SA07 (asthma).

Figure 6.2: Correlation of current RR with past PM-values (both smoothed).



(a) All subjects.

(b) Per subject.

Figure 6.3: Relation between LSTM TS model 1-minute ahead prediction errors and air pollution 11 minutes earlier: scatter plots with linear regression line.

Chapter 7

Conclusion

The aims of this dissertation were twofold: firstly, the APCAPS pilot data was analysed in detail to confirm the value of the data collection using the RESpeck and the validity of the results. Secondly, the same RESpeck data was used to investigate if a person's respiratory rate can be predicted ahead of time. It was also investigated how air pollution relates to respiration in people with asthma and COPD, and whether this could be utilised to improve the predictions.

The first goal has been achieved to a large extent. A number of spatiotemporal attributes can be inferred from the RESpeck data, containing information on the subjects'activities, location, and sleep patterns, besides the physiological breathing and physical activity data. The results of analysis performed on the extracted attributes are in line with published results. The only exceptions were the location-based attributes, which, although accurate themselves, could not be used as valid proxies for lifestyle.

The second goal of predicting the respiratory rate has also achieved substantial results and to the best of the author's knowledge has not been attempted before, making it a promising avenue for further research. Using a neural network with time series and fixed inputs, the respiratory rate could be predicted directly with an MAE of around 1.5 BPM for one minute ahead, rising up to 2.7 BPM for ten minutes ahead. Although these evaluation metrics were roughly equal to the baseline, sample plots of predictions for individual subjects showed that the predictions matched the approximate direction of change much better.

The alternative approach in which models were estimated using the respiratory trend rather than the respiratory rate as outcome, is also encouraging. Taking the trend simplified the problem substantially, which translated into a much lower MAE for all the models. The best estimated model, using time series, first and second differences, and fixed features as inputs, managed to predict one minute ahead with an MAE of only 0.4 BPM, increasing to around 1.75 BPM for a prediction window of ten minutes. Similar to the direct respiratory rate predictions, this was close to the best baseline model when measured by the evaluation metrics, but appeared to be slightly better after visual inspection of predictions plotted onto the true trend. For any future work on the topic, it is therefore highly recommended to devise a more advanced evaluation metric that describes the actual utility of a model better than looking at the MAE, RMSE, or percentage of predictions with the correct direction.

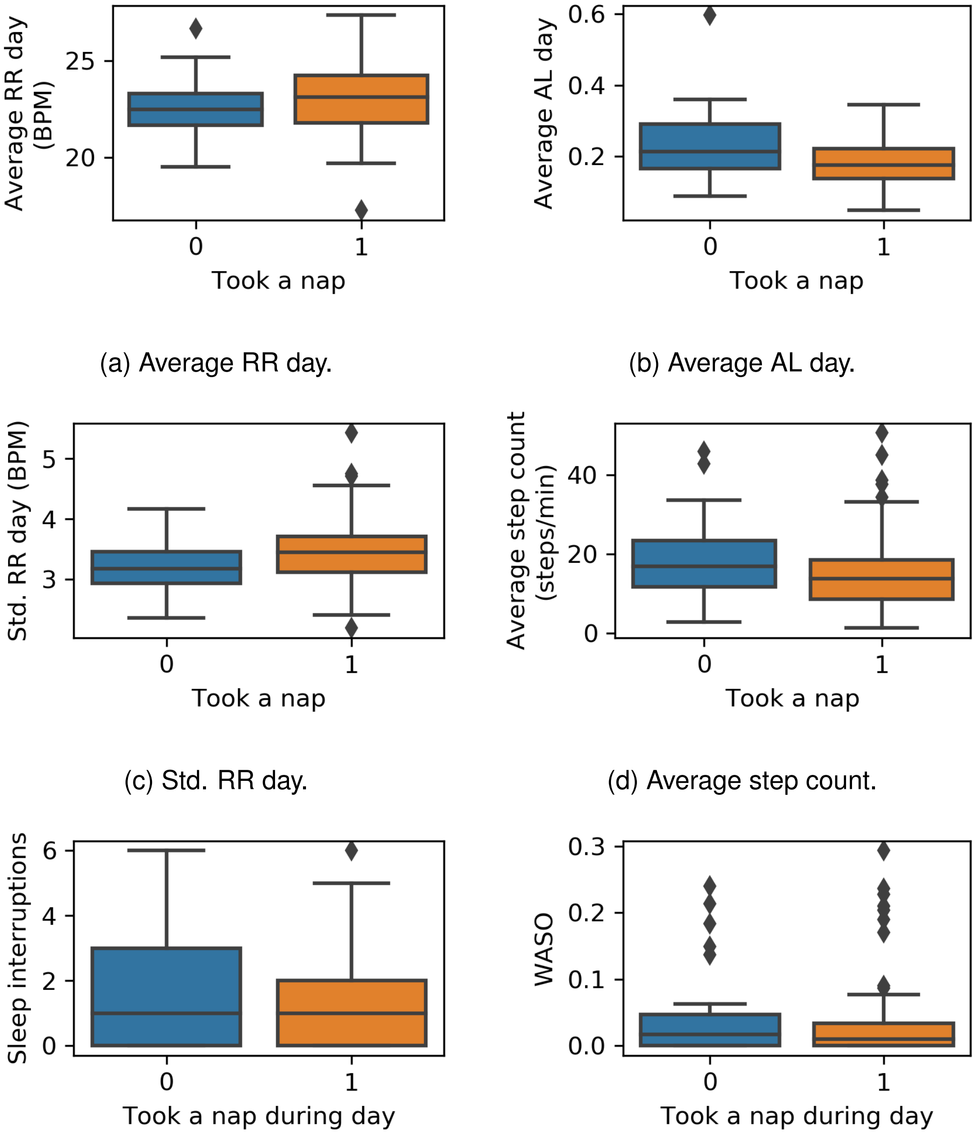
For both types of predictions alike, two observations were made that could contribute to improving the predictions further. The first is that sudden changes in the respiratory rate were responsible for most of the large prediction errors. It is therefore advisable to look into what causes such abrupt changes and incorporate this in the models, if possible. This is especially important with potential medical applications in mind. A hypothesis stemming from the analysis that was carried out already is that the sleep cycle plays a role in this. The second observation is that the average prediction error is highly variable between subjects and the decisive factor for this is the variance of the respiratory rate. It may therefore be worthwhile to collect extra data on subjects with a high-variance respiratory rate, or to oversample such subjects when estimating the models.

Besides these suggestions, another major direction for future research is the relationship between air pollution and respiration, both in the context of respiratory rate prediction and general health. The analysis done on the TackSHS data confirms that these two variables are highly correlated for people with asthma or COPD, even when correcting for confounders such as the activity type and past respiratory rates. Combining this with the observation that applying the APCAPS-estimated respiration models to the TackSHS subjects leads to predictions that are too low on average when the air pollution was higher, it seems beneficial to include air pollution in future respiratory rate predictions. It was also found that the exposure-response relation differed a lot between the people in this (small) sample. Future analysis within the TackSHS study should be able to provide some insights into the determinants of certain exposure-response patterns.

Overall, the work carried out in this dissertation provides a good foundation for several directions of future work. The analysis on the APCAPS pilot data encourages the continuation of the proposed study. The work on respiratory rate predictions has resulted in basic but already useful models, as well as suggestions for improvement. Ultimately, the goal of this would be to build a medical application for monitoring the respiratory rate. Lastly, the exploration of air pollution data has given an outlook on the possibilities of the TackSHS project of predicting changes in respiratory rate at different levels of air pollution.

Appendix A

Boxplots of significant subgroup differences



(a) Average RR day.

(b) Average AL day.

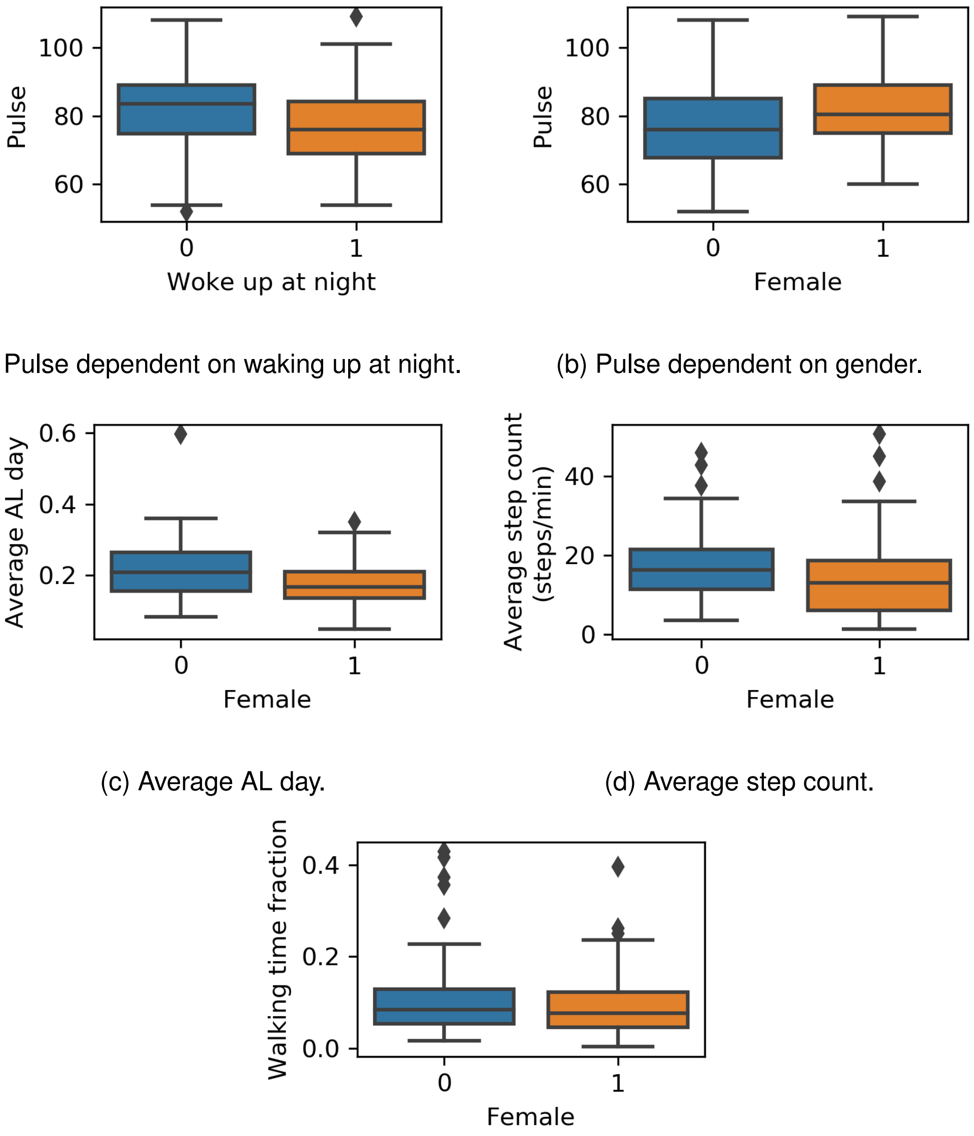
(c) Std. RR day.

(d) Average step count.

(e) Sleep interruptions.

(f) WASO.

Figure A.1: Boxplots of significant subgroup differences: napping behaviour.



(a) Pulse dependent on waking up at night.

(b) Pulse dependent on gender.

(c) Average AL day.

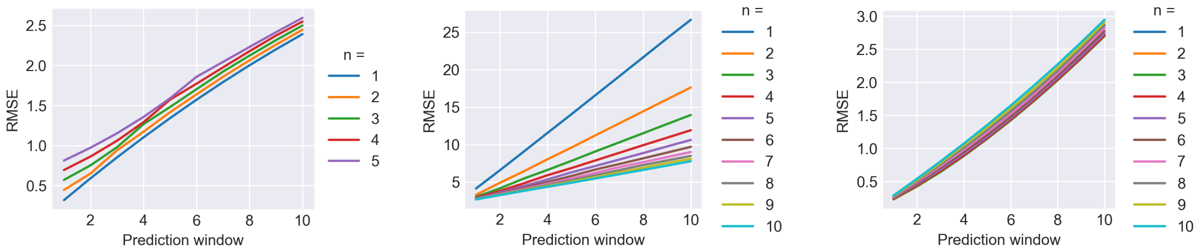
(d) Average step count.

(e) Walking time fraction.

Figure A.2: Boxplots of significant subgroup differences: waking at night and gender.

Appendix B

Plots used for hyperparameter tuning



(a) Mean for trend prediction.

(b) Slope for RR prediction.

(c) Slope for trend prediction.

Figure B.1: RMSE of n-period models against prediction window as a function of n.

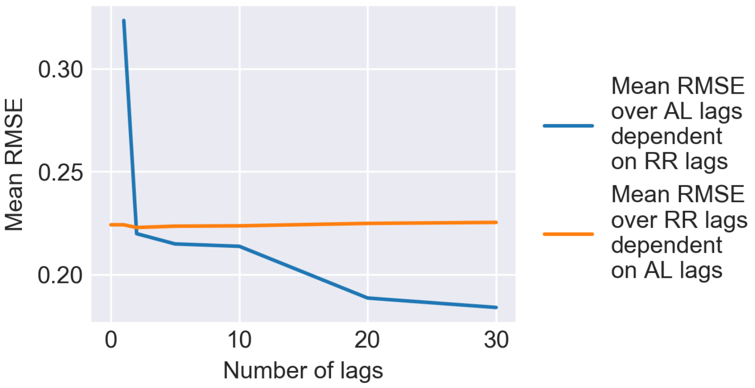
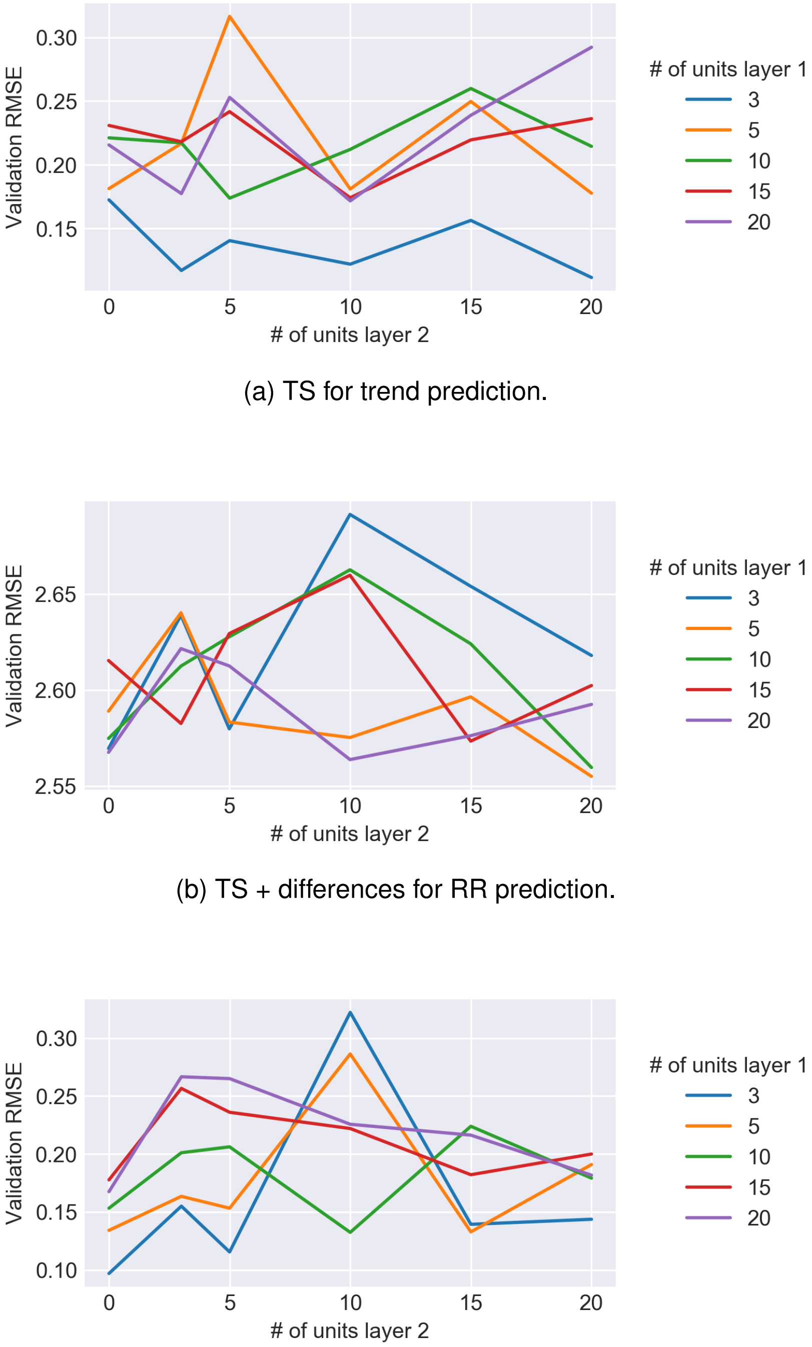


Figure B.2: Mean test RMSE over AL/RR lags for each RR/AL lag option, Ridge regression for trend prediction.

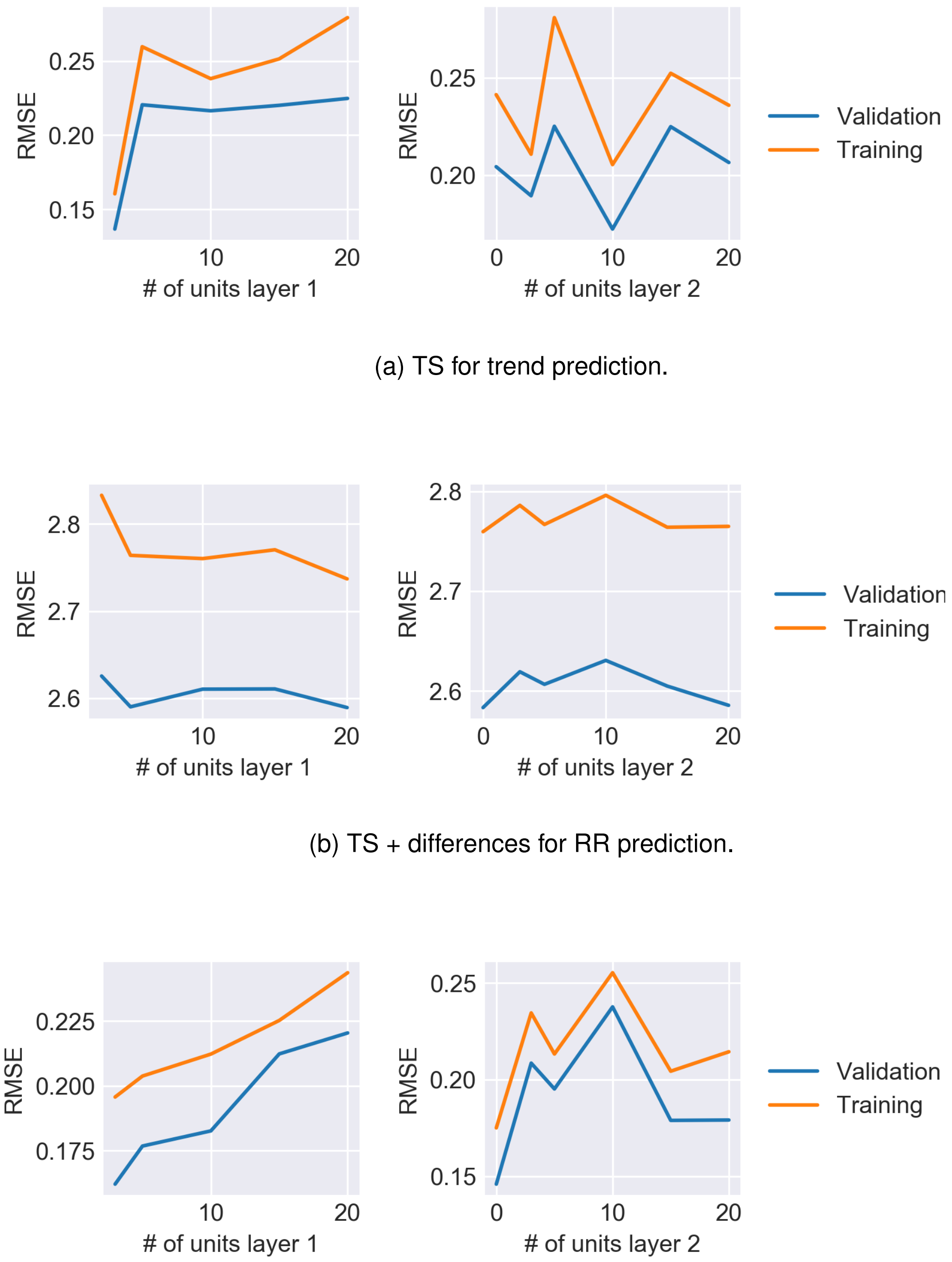


(a) TS for trend prediction.

(b) TS + differences for RR prediction.

(c) TS + differences for trend prediction.

Figure B.3: Validation RMSE depending on number of units in LSTM layer 1 and 2 for models with TS inputs only.

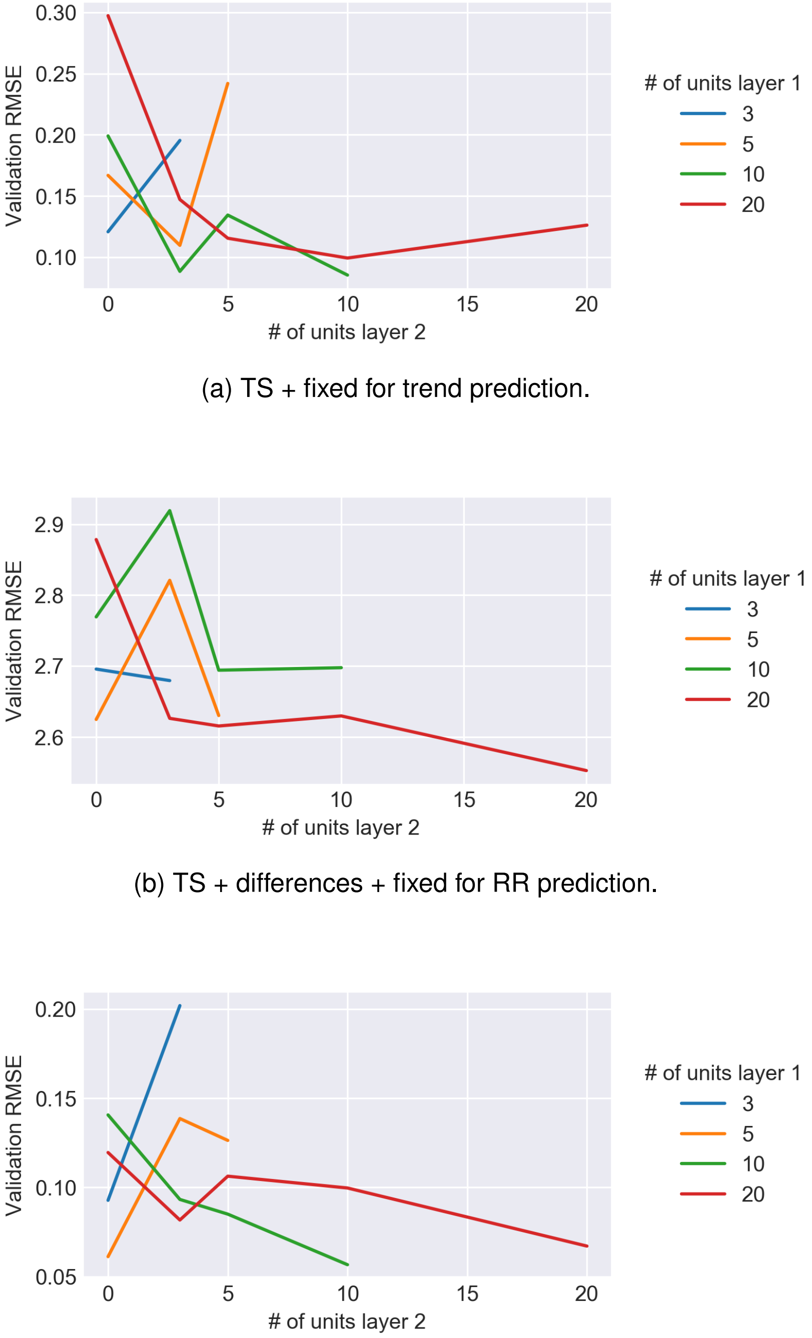


(a) TS for trend prediction.

(b) TS + differences for RR prediction.

(c) TS + differences for trend prediction.

Figure B.4: Mean RMSE over number of units in layer 2 (left) and layer 1 (right) for models with TS inputs only.

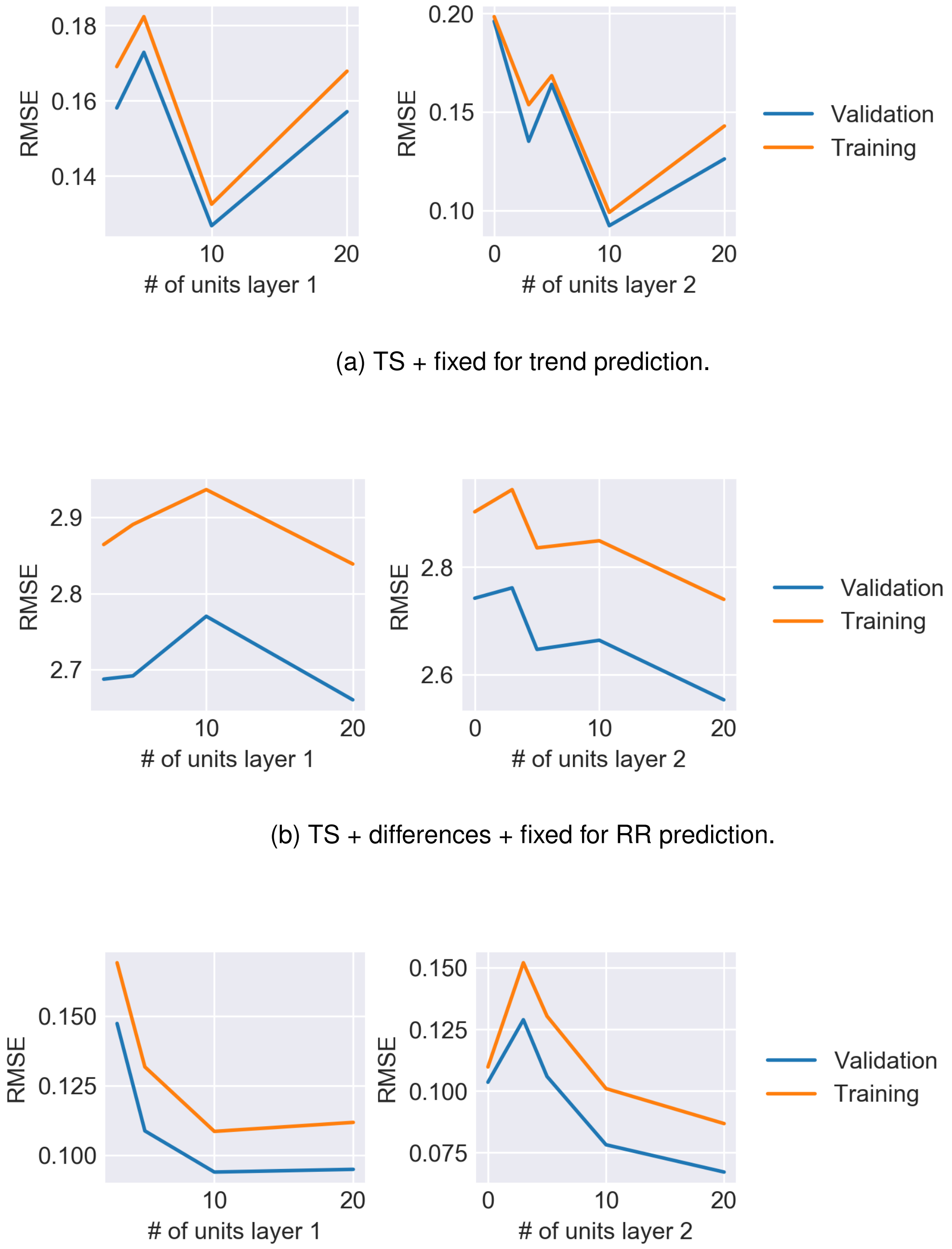


(a) TS + fixed for trend prediction.

(b) TS + differences + fixed for RR prediction.

(c) TS + differences + fixed for trend prediction.

Figure B.5: Validation RMSE depending on number of units in LSTM layer 1 and 2 for models with TS and fixed inputs.



(a) TS + fixed for trend prediction.

(b) TS + differences + fixed for RR prediction.

(c) TS + differences + fixed for trend prediction.

Figure B.6: Mean RMSE over number of units in layer 2 (left) and layer 1 (right) for models with TS and fixed inputs.

Bibliography

Alfaris, N., Wadden, T. A., Sarwer, D. B., Diwald, L., Volger, S., Hong, P., Baxely, A., Minnick, A. M., Vetter, M. L., Berkowitz, R. I., and Chittams, J. (2015). Effects of a 2-year behavioral weight loss intervention on sleep and mood in obese individuals treated in primary care practice. *Obesity*, 23(3):558–564.

Argerich, S., Herrera, S., Benito, S., and Giraldo, B. F. (2016). Evaluation of periodic breathing in respiratory flow signal of elderly patients using SVM and linear discriminant analysis. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pages 4276–4279.

Arvind, D. K., Mann, J., Bates, A., and Kotsev, K. (2016). The AirSpeck Family of Static and Mobile Wireless Air Quality Monitors. In *Proceedings - 19th Euromicro Conference on Digital System Design, DSD 2016*, pages 207–214, Cyprus.

Awad, K. M., Malhotra, A., Barnet, J. H., Quan, S. F., and Peppard, P. E. (2012). Exercise is associated with a reduced incidence of sleep-disordered breathing. *American Journal of Medicine*, 125(5):485–490.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning Long-Term Dependencies with Gradient Descent is Difficult. *IEEE Transaction on Neural Networks*, 5(2):157–166.

Bianchi, F. M., Maiorino, E., Kampffmeyer, M. C., Rizzi, A., and Jenssen, R. (2017). *Recurrent Neural Networks for Short-Term Load Forecasting: An Overview and Comparative Analysis*. Springer International Publishing, Cham.

Bontempi, G., Ben Taieb, S., and Le Borgne, Y. A. (2013). Machine Learning Strategies for Time Series Forecasting. In Aufaure, M. and Zimányi, E., editors, *Business Intelligence. eBISS 2012. Lecture Notes in Business Information Processing*, pages 62–77. Springer-Verlag, Berlin, Heidelberg.

Brower, K. J. and Hall, J. M. (2001). Effects of Age and Alcoholism on Sleep: A Controlled Study. *Journal of Studies on Alcohol*, 62(3):335–343.

Christensen, M. A., Bettencourt, L., Kaye, L., Moturu, S. T., Nguyen, K. T., Olgin, J. E., Pletcher, M. J., and Marcus, G. M. (2016). Direct measurements of smartphone screen-time: Relationships with demographics and sleep. *PLoS ONE*, 11(11):1–14.

Conroy, D. E., Elavsky, S., Doerksen, S. E., and Maher, J. P. (2013). A Daily Process Analysis of Intentions and Physical Activity in College Students. *J Sport Exerc Psychol.*, 35(5):493–502.

D'Agostino, R. B. (1971). An Omnibus Test of Normality for Moderate and Large Size Samples. *Biometrika*, 58(2):341–348.

D'Agostino, R. B. and Pearson, E. S. (1973). Tests for Departure from Normality. Empirical Results for the Distributions of b2 and b1. *Biometrika*, 60(3):613–622.

Drummond, G. B., Bates, A., Mann, J., and Arvind, D. K. (2013). Characterization of breathing patterns during patient-controlled opioid analgesia. *British Journal of Anaesthesia*, 111(6):971–978.

Ebrahim, I. O., Shapiro, C. M., Williams, A. J., and Fenwick, P. B. (2013). Alcohol and Sleep I: Effects on Normal Sleep. *Alcoholism: Clinical and Experimental Research*, 37(4):539–549.

Gonnissen, H. K., Adam, T. C., Hursel, R., Rutters, F., Verhoef, S. P., and Westerterp-Plantenga, M. S. (2013). Sleep duration, sleep quality and body weight: Parallel developments. *Physiology and Behavior*, 121(2013):112–116.

Gupta, N. K., Mueller, W. H., Chan, W., and Meininger, J. C. (2002). Is obesity associated with poor sleep quality in adolescents? *American Journal of Human Biology*, 14(6):762–768.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Karimi, S., Soroush, A., Towhidi, F., Makhsosi, B. R., Karimi, M., Jamehshorani, S., Akhgar, A., Fakhri, M., and Abdi, A. (2016). Surveying the effects of an exercise program on the sleep quality of elderly male. pages 997–1002.

Keras (2018a). Keras functional API guide. https://keras.io/getting-started/functional-api-guide/. Accessed 15-08-2018.

Keras (2018b). Keras: The Python Deep Learning library v2.2.0. https://keras.io/. Accessed 15-08-2018.

Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *CoRR*, abs/1412.6.

Kinra, S., Radha Krishna, K. V., Kuper, H., Rameshwar Sarma, K. V., Prabhakaran, P., Gupta, V., Walia, G. K., Bhogadi, S., Kulkarni, B., Kumar, A., Aggarwal, A., Gupta, R., Prabhakaran, D., Srinath Reddy, K., Smith, G. D., Ben-Shlomo, Y., and Ebrahim, S. (2014). Cohort profile: Andhra Pradesh children and parents study (APCAPS). *International Journal of Epidemiology*, 43(5):1417–1424.

McGee, S. (2012). Respiratory Rate and Abnormal Breathing Patterns. In *Evidence-Based Physical Diagnosis*, chapter 18, pages 145–155. Elsevier, Philadelphia, 3rd edition.

Mendelson, M., Borowik, A., Michallet, A. S., Perrin, C., Monneret, D., Faure, P., Levy, P., Pépin, J. L., Wuyam, B., and Flore, P. (2016). Sleep quality, sleep duration and physical activity in obese adolescents: Effects of exercise training. *Pediatric Obesity*, 11(1):26–32.

Mitchell, J. A., Godbole, S., Moran, K., Murray, K., James, P., Laden, F., Hipp, J. A., Kerr, J., and Glanz, K. (2016). No Evidence of Reciprocal Associations between Daily Sleep and Physical Activity. *Medicine and Science in Sports and Exercise*, 48(10):1950–1956.

Morgan, K. (1987). *Sleep and Ageing*. Croom Helm, London.

Nio, A. Q., Stöhr, E. J., and Shave, R. (2015). The female human heart at rest and during exercise: A review. *European Journal of Sport Science*, 15(4):286–295.

Olah, C. (2015). Understanding LSTM Networks. http://colah.github.io/posts/2015-08-Understanding-LSTMs/. Accessed 26-07-2018.

Ortega, F. B., Chillón, P., Ruiz, J. R., Delgado, M., Albers, U., Álvarez-Granda, J. L., Marcos, A., Moreno, L. A., and Castillo, M. J. (2010). Sleep patterns in Spanish adolescents: Associations with TV watching and leisure-time physical activity. *European Journal of Applied Physiology*, 110(3):563–573.

Park, S.-Y., Oh, M.-K., Lee, B.-S., Kim, H.-G., Lee, W.-J., Lee, J.-H., Lim, J.-T., and

Kim, J.-Y. (2015). The Effects of Alcohol on Quality of Sleep. *Korean Journal of Family Medicine*, 36(6):294–9.

Pesonen, A. K., Sjöstén, N. M., Matthews, K. A., Heinonen, K., Martikainen, S., Kajantie, E., Tammelin, T., Eriksson, J. G., Strandberg, T., and Räikkönen, K. (2011). Temporal associations between daytime physical activity and sleep in children. *PLoS ONE*, 6(8):4–9.

Prasertsung, P. and Horanont, T. (2016). A classification of accelerometer data to differentiate pedestrian state. In *2016 International Computer Science and Engineering Conference (ICSEC)*, Chiang Mai.

Research proposal APCAPS (2018). Using Citizen Science to create a Sustainable Urban Health Laboratory at the site of the Andhra Pradesh Children and Parents Study (APCAPS). Technical report, private communication.

Riley, T. L., Ferber, R., Greenberg, R., Hauri, P. J., Howard, G. F., and Orr, W. C. (1985). *Clinical Aspects of Sleep & Sleep Disturbance*. Butterworth Publishers, Boston.

Roveda, E., Sciolla, C., Montaruli, A., Calogiuri, G., Angeli, A., and Carandente, F. (2011). Effects of endurance and strength acute exercise on night sleep quality. *International SportMed Journal*, 12(3):113–124.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.0.

SciPy (2018). Stats Normaltest v1.1.0. https://docs.scipy.org/doc/scipy/reference/generated/-scipy.stats.normaltest.html. Accessed 08-08-2018.

Seefeldt, V., Malina, R. M., and Clark, M. A. (2002). Factors affecting levels of physical activity in adults. *Sports Medicine*, 32(3):143–168.

Segal, L. N., Oei, E., Oppenheimer, B. W., Goldring, R. M., Bustami, R. T., Ruggiero, S., Berger, K. I., and Fiel, S. B. (2010). Evolution of pattern of breathing during a spontaneous breathing trial predicts successful extubation. *Intensive Care Medicine*, 36(3):487–495.

Shanmathi, V. and Jeyanthi, S. (2017). Detection and Classification of Respiration Disorder Based On Breathing Pattern Using Fuzzy Min-Max Classifier. *Advances in natural and applied sciences*, 11(7):433–440.

Sherrill, D. L., Kotchou, K., and Quan, S. F. (1998). Association of Physical Activity and Human Sleep Disorders. *Arch Intern Med*, 158(17):1894–8.

Shrivastava, D., Jung, S., Saadat, M., Sirohi, R., and Crewson, K. (2014). How to interpret the results of a sleep study. *Journal of Community Hospital Internal Medicine Perspectives*, 4(5):24983.

Subramanian, S., Hesselbacher, S., Mattewal, A., and Surani, S. (2013). Gender and age influence the effects of slow-wave sleep on respiration in patients with obstructive sleep apnea. *Sleep and Breathing*, 17(1):51–56.

TensorFlow (2018). TensorFlow v1.8.0. https://www.tensorflow.org/. Accessed 15-08-2018.

Várady, P., Micsik, T., Benedek, S., and Benyó, Z. (2002). A novel method for the detection of apnea and hypopnea events in respiration signals. *IEEE Transactions on Biomedical Engineering*, 49(9):936–942.

Ward, N. S. and Levy, M. M. (2017). *Sepsis: Definitions, Pathophysiology and the Challenge of Beside Management*. Springer International Publishing AG, Cham.

World Health Organization (2017). Improving the prevention, diagnosis and clinical management of sepsis. Technical Report January, World Health Organization.

Youngstedt, S. D., O'Connor, P. J., and Dishman, R. K. (1996). From wake to sleep. The effects of acute exercise on sleep: a quantitative synthesis. *American sleep disorders association and sleep research society*, 20(3):203–214.