Aalto University
School of Economics

A Statistical Model of Disability Pension Risk

Quantitative Methods of Economics

Master's thesis

Mikhail Savin

2010

Department of Business Technology
Aalto University
School of Economics

Abstract

The aim of this study was to confirm the existence of and explore a hypothesized statistical relationship between sickness absence data and disability pensions within a population on an individual level. Using this information a statistical model was built to forecast disability pension risk and to study the distribution of employee health within organizations. The development of this model was motivated by the opportunities it would provide in financial forecasting and employee rehabilitation in Finnish government offices and agencies.

Within the scope of this thesis research on absenteeism and early retirement was reviewed. The reviewed literature covered a variety of geographical areas and incorporated several different approaches to the analysis of phenomena under study. The role of behavioral and psychological factors in these decisions was stressed and this was also the focus of the literature review. The theoretical insights were then used in the explorative analysis of a personal-level data set provided by the Finnish State Treasury and Ministry of Finance. The main model which was developed in this study was a state space model with logistic transfer functions. The specification of the states was performed on a theoretical basis, while the transfer functions were estimated statistically.

The findings of this study can be separated into two areas – academic findings related to sickness absences and the developed model for practical use. The exploratory data analysis has allowed making several important observations concerning sickness absence patterns prior to disability pension events. Two distinct sickness absence patterns were identified. Each of the sickness absence patterns has specific parameters in terms of duration and quantity of sickness absences.

The practical result of the study is the development of the state space model for evaluation of disability pension risk. The model provides reasonable short term forecasting power and allows studying and comparing employee health distributions within and between organizations. In this way the model acts as a powerful financial and managerial tool.

Key words: disability pensions, sickness absences, state space, early retirement, absenteeism

Table of Contents

List of Figures

List of Tables

1 Introduction

Employee welfare is a core value for a well-functioning organization. It drives work motivation and stimulates performance. An effective social support and pension system is one of the key requirements for the achievements of high level of employee and social welfare. Development of such social support systems requires a high level of understanding of principles and underlying dynamics of employee health and wellbeing as well as statistical models, which would allow effective forecasting of both individual and aggregate trends in these areas. This thesis focuses on quantitative analysis of disability pension risk and continues the model development presented in "Sickness Absences as an Indicator of Disability Pension Risk" (Savin, 2009) by extending the model to the individual level. As an introduction to this study, I present a more precise overview of research motivation and further information about the methodology used in the scope of this thesis.

1.1 Motivation for research of pension dynamics

The public pension security and employee disability compensations form a fundamental aspect of the system which guarantees the employee wellbeing. Work ability diminishes with age and as a result of work-related physical or psychological difficulties, for example, accidents or illness. The support and rehabilitation of the employees susceptible to the mentioned work disability threats is of key importance, especially as the age structure of the working population changes. In addition to the rehabilitation possibilities, this part of the welfare system is becoming a central source of expenditure for governmental organizations and the risks become especially noticeable in the budgeting procedures of smaller departments, where a single disability pension may have a dramatic effect on the financials. As a result, there is an increasing demand for accurate forecasting information related to the employee work ability risks, which could be used for both financial risk assessment and as a trigger for introduction of early intervention programs in an organization.

The opportunities offered by such forecasting models are especially extensive in the public sector, where the employee information is systematically collected and stored for further analysis. The availability of a wide range of systematic information concerning employee characteristics, health state and work ability provides an opportunity to explore the

relationship between the different categories of data and construct a forecasting model based on the observed patterns. The forecasts then could be shared to coordinate both the pre-emptive employee support by the health-care service providers and the financial risk levels could be provided by the State insurance institution in order to simplify the risk management and financial accounting processes in the government offices and agencies.

As discussed in Savin (2009), the current research in the area of disability pensions and work absenteeism is, on the other hand, highly fragmented and mostly focused on the analysis of long-term effects from medical and psychological viewpoints. There is an abundance of literature focusing on each of these issues separately, but only a relatively small share of the literature establishes the link between these and none offers a quantitative model to describe this relationship. **Table 1** from Savin (2009) offers a good summary of the available literature in the specified areas.

Table 1. Literature query results in several databases in 2009

	Search Query				
Database	**Sickness**	**Disability**	**Sickness absence**	**Disability pension**	**Disability pension AND Sickness absence**
EBSCO EconLit	243	1573	19	36	0
Emerald MCB University Press journals	1924	2444	907	236	0 (journals) / 5 (books)
HSE Helecon MIX	7	26	1	2	0
HSE Helecon SCIMA	45	61	13	7	0

Due to the importance of the topic for public welfare provision and due to the lack of powerful quantitative models to forecast specific types of pension expenses, the Finnish State Treasury has expressed interest in the development of this model, acting as an initiator for the original research in this area. The representatives from the Finnish State Treasury have expressed view that a more precise risk predictor for the financial expenses related to disability pensions would most likely be greatly appreciated by their customers in the form of the government offices and agencies. In addition to the improvement in financial forecasting, a model of disability pensions could also provide a valuable opportunity for pre-emptive action on an aggregate level without creating concerns with regards to privacy protection.

1.2 Research question

In the light of the previously described deficiencies in the current research and due to the wide range of opportunities the research question selected for this academic study is focused on disability pension forecasting. More concretely, the goal is to analyze individual level sickness absence data for a fairly large population and to develop a forecasting model based on a state-space system, which was analyzed and described in the previous study.

To develop an individual level system we will additionally need to answer questions related to the individual perspective on early retirement. The reason for this is that the relationships modeled within the system are a result of choices made by individual employees. This means that we will have to understand the possible reasoning and motivation for early retirement and search for proxy variables, which would allow quantitative modeling of these aspects. Due to this reason, an additional focus of the literature review will be to augment the review presented in Savin (2009) and to extend it into the area of organizational, behavioral and psychological research. This additional insight will contribute to the development of the set of factors, which will be included into the state space model developed in this thesis.

1.3 Research methodology

The thesis is organized into three distinct parts with specific methodology used in each one of them. The starting point of the research, constituting the first part of the thesis (*Section 2*), is a literature review, where previous research is augmented with recent quantitative studies and academic research in the areas of organizational behavior and psychology. The observations are then linked to the quantitative model by analyzing and creating various factors and indicators, which will be developed and verified in the following parts of the thesis.

In the second part of the thesis (*Section 3*) the focus is placed on the quantitative analysis of the data and of the population under study. The data, which represents the employees working in government offices and agencies, has been collected from two distinct sources and it has to be checked for integrity and analyzed. In addition to this, different types of disability pensions will be considered and a hypothesis for the relationship between disability pension type and the corresponding sickness absence pattern will be created. The main idea behind this population analysis is to understand the specifics of the data better and use this knowledge in

model development. This will allow us to optimize the model in the light of these specifics and improve its forecasting power.

In the next part of the thesis (*Section 4*), the focus will be placed on the development of a quantitative model to forecast the likelihood of individuals moving to disability pension. The model will be developed according to general principles described in Savin (2009) and will be fit and tested on the data set presented in the second part of the thesis. The testing of the model will take place by comparing the state space model to a more simple logistic regression model. The predictive power of the state space model will be evaluated and the trade-off between complexity of the model and the predictive power will be discussed.

Dwelling on the results of the analysis in the final section of the report (*Section 5*), I will evaluate the model performance and present a view on the application perspectives for the developed model as well as demonstrate a methodology and a visual indicator which can be applied in organizations using the model. In addition to this, in the conclusion of this thesis, directions and implications for future research will be provided.

Section summary

The study of disability pensions and their forecasting are of growing importance due to demographic changes in the population, which stress the role of employee rehabilitation and prolongation of working life.

This study extends the BSc thesis by Savin (2009) by including theoretical discussion of behavioral and psychological factors and by developing a model on personal-level data.

2 Review of Academic Literature

In this section I provide an overview of literature and research related to the employee sickness absences and disability pensions. In the first part of this section I briefly review the research findings from Savin (2009) and provide some additional insight into the relationship between sickness absence and disability pension relationships. The second part of the literature review focuses on the psychological and behavioral factors associated with these phenomena, while in the last part of this section the statistical modeling techniques applicable to this thesis are reviewed.

2.1 Employee sickness absences and disability pensions

The lack of literature linking the phenomena of disability pensions to the absenteeism behavior of the employees has been underlined in Savin (2009). To confirm the previous findings and as a part of the effort to evaluate the new literature in the related fields, the analysis of article search results was performed again within this thesis work. **Table 2** presents the results of the repeated search. Additionally, the differences with regards to the previous findings are presented. The less relevant databases have been dropped and OvidSP has been added to demonstrate the volume of psychological and behavioral research in this field.

Table 2. Reviewed literature query results from several databases in 2010

	Search Query				
Database	**Sickness**	**Disability**	**Sickness absence**	**Disability pension**	**Disability pension AND Sickness absence**
EBSCO EconLit *(general)*	261 (+18)	1680 (+107)	25 (+4)	39 (+3)	0 (+0)
Emerald MCB University Press journals *(general)*	1994 (+70)	2598 (+154)	944 (+37)	253 (+17)	0 (journals) / 5 (books) (+0)
OvidSP *(behavioral sciences)*	46690	10709	1704	960	4 (none relevant)

The query results show a fairly large increase in the volume of research in the fields related to sickness absences and disability pensions that has been released within a timeframe of less

than 1 year. This stresses the relevance of these fields of research for the challenges and population dynamics faced by the society. Further, the volume of psychological and behavioral research on disability pensions and sickness absences indicates that the psychological aspects of these phenomena are significantly more researched than the economic aspects. This also indicates the fact that augmenting the model developed in Savin (2009) with the application of qualitative frameworks developed in psychological and behavioral research may provide a more complete picture of phenomena under analysis. It is also important to note that the niche of academic studies addressed by this paper, namely the link between sickness absence patterns and disability pensions, remains poorly researched. There is a very low amount of academic literature in this area and no increase in its amount has been noticed during the past year.

Due to the noticeable increase in the amount of academic research in the areas closely related to this study (i.e. research on sickness absences and disability pensions in separate) this section will cover the relevant literature and will provide a review of the findings presented in Savin (2009). Initially, the literature concerning sickness absences and disability pensions will be analyzed separately and then several papers linking these two concepts will be presented.

2.1.1 Sickness absences

Business literature focuses mostly on the human resource management perspective on sickness absences. From this perspective, the sickness absences are viewed as not only costly occasions for an enterprise, but their low predictability also leads to disruptions in work processes and decreases efficiency as the reduced labor force has to be reshuffled to fill the created gap (Muir, 1983). This makes the reduction of sickness absences and their increased predictability a core focus for business-related research on sickness absences in workplaces.

The role of sickness absences as an important cost for businesses has been increasing for a long period of time. For example, from 1981 to 1991 the rate of sickness absences in some areas of UK has doubled (Muir, 1994) with 35% of these absences reported to be related to work stress. Especially due to the fact that in many cases the national legislation forbids the employer to require doctor notification for short period sickness absences, Muir (1994) underlines the fact that a large part of the sickness absence volume is just motivation-related absenteeism in disguise. From the business perspective, distinguishing between these types of sickness absences is crucial and within the scope of this thesis there is a logical reason to believe that if the underlying factors related to sickness absences have different nature, their

relationship with disability pensions will differ. From the business perspective, the method to control these sickness absences suggested by Muir (1983, 1994) is directly addressing the employees. In his articles Muir suggests that, statistically, many employees reduce their number of sickness absences in response to direct contact from the management, indicating the fact that the absences are caused by light sicknesses or by completely unrelated factors.

Empirical studies provide further support for the role of factors unrelated to health determining the rate of sickness absences at workplaces. For example, the characteristics of the employment relationships and the characteristics of the family of the employee have a fairly strong correlation with the number of sickness absences (Toivanen et al. 2008). Many of the factors in the study performed by Toivanen et al. (2008) were especially important for the blue-collar and lower white-collar employees. Nevertheless, for white-collar employees the number of children below age 7 was a strong predictor of increased absence rates. These natural family-related determinants of sickness absences suggest that within the scope of this study, special attention should be paid to the nature of sickness absences under analysis and un-verified short-term absences may not be significant determinants of disability pensions. Nevertheless, these family-work conflict, stress- and motivation-related absences could be valuable indicators of psychological work disability and the resulting disability pensions.

The value of the records associated with the employee sickness absences is also underlined by Muir (1983), where he briefly states that the data could be used to evaluate differences between different enterprise divisions or employee positions with the aim to discover possible underlying factors in the employment relationship or job description which need to be improved to avoid both health deterioration and to increase the motivation of the employees to attend the workplace.

The analysis of academic literature related to sickness absences establishes the key roles of underlying reasons for sickness absences for the model developed in this paper. Howarth (2005) summarizes the key absence determinants to be medical, stress-related, motivational, domestic, unavoidable and planned. Most of these factors could result in sickness absences, especially if the organization maintains a strict policy towards other absence types. Due to the fact that short-term sickness absences are separated from the longer absences, which must be verified by a doctor, the differences between their determinants will be considered in the model development.

2.1.2 Disability pensions

Literature related to the phenomenon of disability pensions takes a focus on the macroeconomic implications of work disability. In this way disability pensions are presented as factors which affect the total working population of a given country and the dynamics of disability pension rates play an important role in determining the age structure of this working population. A wide range of quantitative studies addressing the issue of disability pensions from this perspective is available. Within the scope of this thesis, studies related to the Nordic region will be of interest, because they would allow illustrating the working population dynamics in regions with similar population structure and legislation related to retirement as the region under study. One study of this type was reviewed previously in Savin (2009) and several more will be presented in this section.

In the majority of papers under analysis, disability pensions are analyzed as one of the channels for early retirement from workplaces. Empirical evidence shows that the rate of early retirement has been increasing during the recent years (Bolin et al. 2008). Being the largest cause of retirement unrelated to age (Savin, 2009), disability pensions are the focal aspect of studies on early retirement. Nevertheless, it is important to note that the increased rates of early retirement cannot be explained fully by changes in the disability pension levels (Bolin et al. 2008). Additionally, disability pensions tend to be much more probable in the period when the employee is already within several years of the age required for old age pension. **Figure 1** illustrates the predicted occurrence of various pension types in Sweden predicted in the study by Bolin et al. (2008) and shows the significant increase in predicted level of disability retirement at age 61-64. At age 65 no new disability pensions are started, because the individuals are transferred to old age pension instead of disability pension.

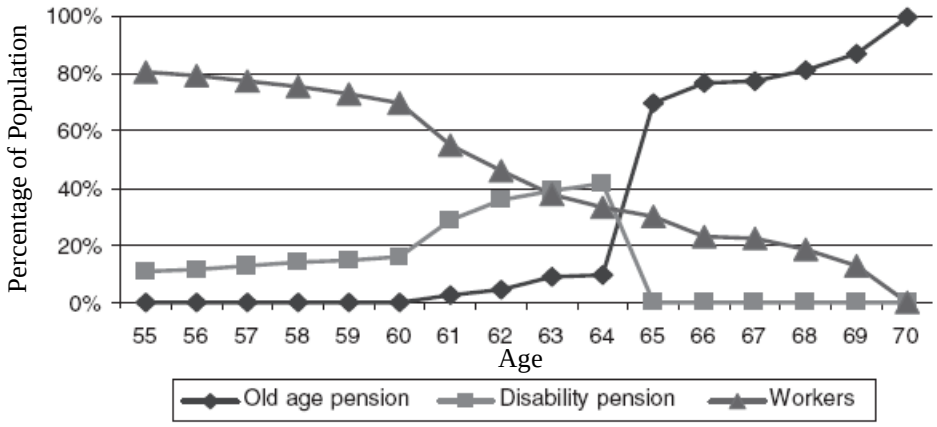

Figure 1. Shares in old age pension, disability pension, and work, by age (one of several alternative scenarios) for years 2010–2040 in Sweden (Source: Bolin et al. 2008)

Bolin et al. (2008) have used the Swedish Ministry of Finance SESIM micro-simulation model to develop their own simulation of the population characteristics in terms of sickness absence, retirement, mobility and a variety of other factors. Due to the number of factors involved in the model and the focus of research on the Swedish pension model the results are mostly related to the dynamics of different pension types in Sweden. Nevertheless, similarly to the situation with sickness absenteeism, the authors state that the distinction between early retirement due to health problems, voluntary early retirement and retirement due to old age is very small, especially in the older age groups. This means that a large proportion of the older population may be eligible for disability pension, while the ones leaving for disability pensions may do this because of economic or motivational factors.

A similar working lifetime study was performed by Nurminen et al. (2005) for the Finnish population. Within the study the authors modeled and simulated the structure of the working population of Finland. As a part of their model, they analyzed the disability pension occurrences for the target population. **Figure 2** presents the observed probability of work disability within the Finnish population. The figure also illustrates the historical dynamics of retirement due to work disability.

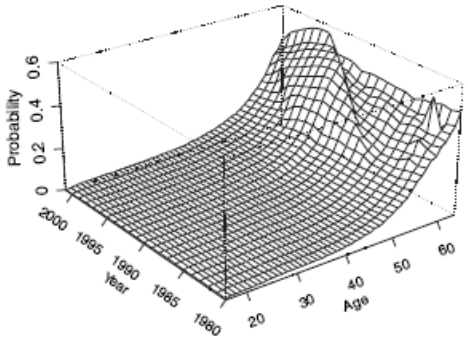

Figure 2. Observed probability of work disability in Finland (Source: Nurminen et al. 2005)

Nurminen et al. (2005) have also tried to create a macro-level model to explain the tendencies in the changing structure of the working population. The approach used was very similar to that of Savin (2009), where a Markov model is used as a method to describe the relationship under analysis. The different states used in the model relate to the possible states of an individual in the population (employed, disabled, dead selected as states in Nurminen et al. 2005).

The brief analysis of core academic literature related to disability has shown that research in this area is scarce and mostly focused on the macroeconomic analysis of the key trends in population structure. Disability pensions are in this way recognized as mechanisms transferring individuals out of the working population and the economic and motivational incentives related to disability pension plans are considered to be an important factor affecting the rate of disability pensions in a population. The study by Nurminen et al. (2005) also confirms the effectiveness of use of Markov chain models in the analysis of pensions and related phenomena within populations.

2.1.3 Statistical relationship between sickness absences and disability pensions

The analysis and modeling of the relationship between sickness absences and disability pensions forms the core of this paper. As previously mentioned, the amount of literature linking these two phenomena is fairly low, but the several pieces of research to be analyzed in this part form an extremely important basis for model development and provide insight into some problems associated with the modeling of these phenomena. Similarly to the case of literature on disability pensions, the analysis of its relationship with sickness absence is predominantly performed on a macro level, where the whole population is analyzed.

Nevertheless, as stated by Savin (2009), the analysis of the population may be highly beneficial in the construction of a model on an individual employee level.

A study by Bratberg, Sturla and Mæland (2008) describes the relationship between the sickness absences with psychiatric diagnoses and the disability pensions which follow and was used in Savin (2009) as an initial basis for supporting the hypothesized link between the two phenomena. The study is limited to sickness related to psychiatric disorders, but it provides a simple framework for the analysis of the link between psychological health problems and disability pensions. As it can be seen from Figure 3, the authors indicate that the disability pension risk is positively correlated with the deprivation index (indicating social separation and lack of psychological wellbeing) for different counties (regions within the country) in Norway. This result is then combined with the observation that the level of deprivation also affects the level of sickness absences related to psychological illness. As a result, the authors state that especially in case of long-term psychological sickness absences the level of deprivation will be an important factor in determining if the employee will be able to return to work or if he or she will move to disability pension. This hypothesized relationship underlines the fact that mediating variables can be used in a statistical model to model the link between sickness absences and disability pensions, which is a part of the modeling process performed within the scope of the thesis.

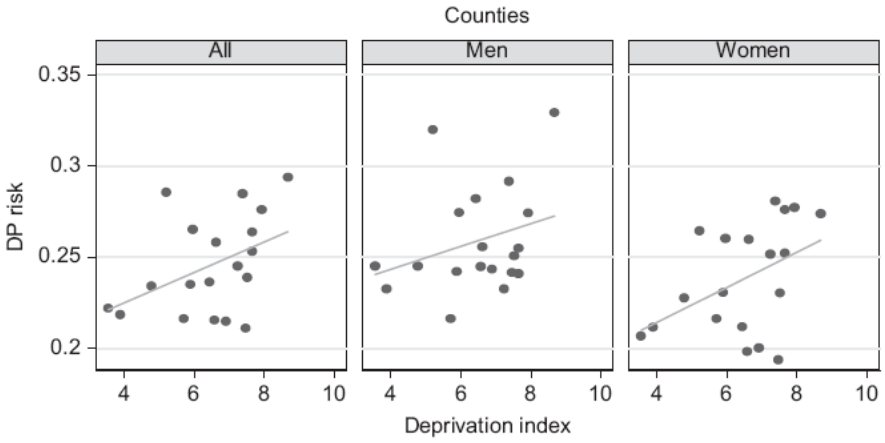

Figure 3. Disability pension risk and deprivation index (Source: Bratberg et al. 2008)

The ability to use other factors, such as the level of deprivation which was used by Bratber et al. (2008), as mediating variables in the model describing disability pension risk is supported by other empirical research in this area. Friis, Ekholm and Hundrup (2008) have performed a study where the relationship between lifestyle, working environment, socio-demographic

factors and the disability pension risk was analyzed in a population of elderly nurses in Denmark. The authors discovered that not only the health- and work-environment related variables were important descriptors, but also socioeconomic factors such as income level and marital status had an important effect on disability pension risk. The nature of this relationship is most likely twofold. Firstly, it can be argued that the socioeconomic factors determine the lifestyle and attitudes of the individuals which in turn are reflected in the health and behavior. On the other hand, factors such as marital status and level of income may affect the motivational constructs and work-life balance desired by the individual. As a result, they may greatly contribute to describing the likelihood of voluntary pensions which are registered as disability pensions due to the economical attractiveness of such option.

Socioeconomic and behavioral factors do not only affect the transition from periodic and long-term sickness absences to disability pension, but also the reverse transition from temporary disability pension back to the state of full-time work. A study by Kaiser, Mattsson, Marklund and Wimo (2007) brings up very similar factors to those outlined in Friis et al. (2008). However, socioeconomic variables such as marital status, education and profession type are also accompanied by a variety of more qualitative self-determined variables related to health perceptions and psychological stability, which receive a primary role as determinants of success of rehabilitating treatment. This underlines the importance of including psychological and behavioral factors or proxies for such factors into the model under development.

A study most relevant and similar to the research focus of this paper has been performed by Wallman et al. (2009) where specifically the registered and reported sickness absence record was used as an explanatory factor to describe the disability pensions rates in the Swedish population with a fairly extensive follow-up period of 16 years. The result of the study showed significant relationship between these two variables, verifying the results presented on the population level by Savin (2009). What is interesting and highly useful in the scope of this study, are the variables related to sickness absences and their significance in the resulting relationship.

The variables which were developed in the study by Wallman et al. (2009) to describe the sickness absences were all derived from the Swedish National Social Insurance Agency, which is highly similar to the data set used in the current study. Several variables were developed:

- *Interval between sickness absences* – this variable was defined as the number of non-compensated (by sickness leave) days between spells of sickness absences
- *Number of sickness absences in 1 year (1Jan – 31Dec)*
- *Number of sickness absences in previous 2 years (1Jan – 31Dec)*
- *Pension and sickness absence type (based on compensation)*

The authors have also noticed that the distribution of sickness absences was positively skewed and performed additional analysis on the log-transformed data, which did not affect the results according to the paper. Logistic regression was used as the method to link sickness absence data and additional background variables to the outcome of disability pension. This is also the proposed initial test set-up which will be developed in this thesis. This simple model will be used as a benchmark against which the state space model will be tested and evaluated. In the logistic regression model, all of the previously listed sickness absence-related predictors were highly significant except for the sickness absence and disability pension types, which was quite surprising. The authors stated that the progression and effects of various sickness types were identical according to the data analysis. This assumption could greatly simplify the data analysis and will be evaluated in the present thesis; however this homogeneity of sickness types and their effects will not be taken for granted.

Literature linking sickness absences and disability pensions inherits many of the approaches and challenges associated with academic research in these two areas. Population-level approach remains the core focus of quantitative models in this area, which is also the approach selected in the paper by Savin (2009), which this academic work will extend and build on. The key challenge identified in the academic literature analyzing sickness absences and disability pensions is the partially indistinguishable nature of voluntary and health-driven sickness absences and disability pensions. In this way, these two phenomena are rather similar in nature and can be both described as concrete choices made by individuals, where motivational, economic and social factors are acting in combination. In this way, the analysis

of psychological and behavioral factors associated with these phenomena is a necessary step in describing them and their relationship. This is the focus of the following section.

2.2 Psychological and behavioral factors

As we have shortly discussed previously, psychological and behavioral variables can also play an important role in decisions related to absenteeism and early retirement. Work motivation, work-family conflict and stress can be important factors, may lead especially to short-term sickness absences and retirement at ages close to the limit for the old age pension. Since these phenomena are very closely related to the relationship under study, this section will be dedicated to the analysis of their effect on this relationship. First, the relationship between psychological and behavioral variables with absenteeism will be discussed. Secondly, the relationship of these variables with disability pension risk will be analyzed and finally implications for employee well-being, prevention and rehabilitation will be discussed.

2.2.1 Psychological and behavioral factors affecting sickness absences

Work absenteeism is not defined solely by the health state and work ability of the individual. The reason for this is the fact that usually there are legislative requirements for the employer to allow short-term sickness absences without any need for medical confirmation of sickness. As a result, taking a sickness absence becomes the most simple and economically viable channel for non-health related absenteeism and is in this way determined by more general factors related to employee motivation for work.

In their systematic review of the causes for employee absenteeism Rhodes and Steers (1981) identify two key mechanisms affecting employee absences – motivation for work and ability to work. According to the authors, these two factors are in constant interplay and only their combination allows explaining the full range of causes for absenteeism. The authors further develop a framework, the schematic of which is presented in **Figure 4**.

From the behavioral and psychological perspectives, the framework has several important relationships. Attendance motivation is selected as a key factor, which affects the final employee attendance, while the ability to attend (e.g. health-related) mediates this relationship. In this way, an employee who may not be in fully healthy state may still attend work due to high levels of attendance motivation and only partially fulfill his tasks due to

health reasons. On the other hand, a fully healthy individual may be reluctant to come to work and take short sickness absences even though the individual may be fully healthy.

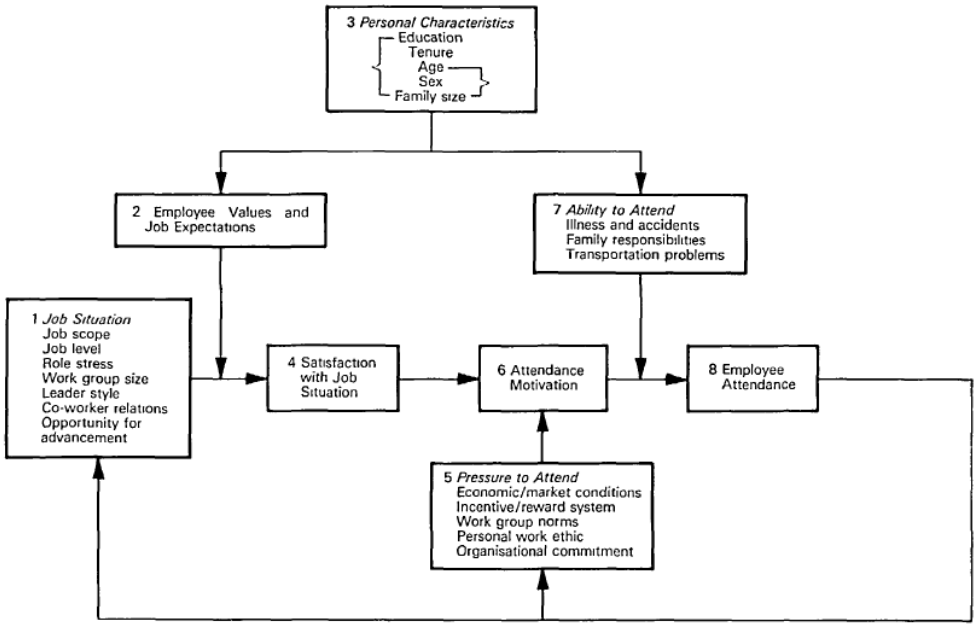

Figure 4. Major influence on employee attendance (Source: Rhodes and Steers, 1981)

In this relationship, our core interest will lie in the understanding of the determinants of attendance motivation and of the mechanism through which the relationship is constructed. According to Rhodes and Steers (1981) the main determinant is the level of job satisfaction, which can be influenced by the set of characteristics related to the job situation: job scope, level, role stress, work group size, leader style, co-worker relations and opportunity for advancement. Additionally, the relationship is mediated by the personal values and expectations of the employee. In this way, the satisfaction is defined by social, professional and personal factors.

Naturally, purely job satisfaction may explain part of the attendance motivation, but it is clear that other motives – especially the economic one, may often be a dominant reason for work attendance. These factors are more related to economic factors described in the previous sections, but also include such determinants as work ethics and commitment. These factors may depend on the cultural and social background of the individual and these background variables could be important descriptors in the model under development.

2.2.2 Psychological and behavioral aspects of work disability

In the understanding of work disability patterns psychological and behavioral factors are no less important than in the case of absenteeism. In this case, there are several possible roles for the behavioral factors. Firstly, they determine the likelihood of disability pensions related to work motivation, stress and other psychological inhibitors of work ability. Additionally, the work environment and the social factors within it determine the level of commitment to the workplace and may affect the likelihood of an individual to actually report mild work disability problem. Finally, these similar constructs affect the ability and the motivation of the individual to return to full-time work in the case of temporary disability pensions. In this way, the types of effects are clearly very similar to those in absenteeism. The existence of some of these effects is supported by academic literature, which will be presented in this section.

It is hard to deny that psychological health is an important determinant of work ability and employee productivity. However, the direction and the nature of this effect is not clear cut – some levels of psychological disability may inhibit work performance, while certain psychological problems, especially if they are treated correctly, can result in almost no work performance differences in comparison to an average employee. A literature overview by Burton, Schulz, Chen and Edington (2008) shows that the type of psychological illness and the level of current treatment determine the level of performance of the employees at their workplace. For example, employees with high levels of depression or suffering from bipolar disorder may attend the workplace normally, but have a significantly lower output. Regular treatment, on the other hand, allows restoring the work ability of such individuals to an acceptable level and avoiding permanent disability pension. Financially, the incremental healthcare costs are justified by the improved productivity (Burton et al., 2008).

In the case of individuals with decreased work ability, for example, due to physical disability, the choice of rehabilitation or simply the choice of not moving to permanent disability pension is defined by both the internal decision of the individual, but is also strongly influenced by the external parameters at the work place. Despite a variety of legislation, the external pressure or simple lack of employment prospects for individuals with physical or psychological work disability is a large problem in rehabilitation and reemployment. In a qualitative study by Newton, Ormerod and Thomas (2007), the authors have found that even physical environment at the workplace is often a barrier for the disabled employees.

The physical environment may produce both a direct physical effect but may also create smaller routine problems which are gradually translated into frustration, lack of motivation and degrading attitude towards the workplace. The study results (Newton et al., 2007) lead us to a conclusion that a concrete employer can create a huge impact on the ability of individuals with mild cases of work disability to stay employed and contribute maximally to the work process. These findings indicate a slight problem related to our model development process, namely the importance of contingent qualitative factors for reemployment of individuals. A thorough analysis of these factors is presented in the theoretical framework developed by James, Cunningham and Dibben (2006). These factors include several social aspects of the workplace and work process design:

- *Support and access to worker representatives* – the structure of the hierarchy at the workplace is extremely important, since it to a large extent determines the motivation and the ability of an individual to report possible psychological or physical problems early on.

- *Availability of specialist advice* – many employees may not be aware of certain aspects of the workplace or their health state which may affect their future disability risk. For this reason, specialist analysis must be available to both managers and employees.

- *Identification and early pre-emptive actions towards vulnerable worker groups* – pre-emptive actions can greatly increase the possibility of quickly rehabilitating the employee and reducing the probability of permanent disability pension. The importance of this factor is also linked to the role of the model developed within this thesis, as it will allow to detect risk-groups and guide pre-emptive actions towards the areas and departments within organizations which require them most.

These factors can not be directly included into the model and finding suitable proxy variables may be very difficult. As stated in Savin (2009), the rehabilitation of individuals is indeed an important process within the proposed model, as it describes movement between two specific states in the state space. Nevertheless, it can also be argued that the analysis of employees of purely government offices and agencies may partially alleviate this problem, because we would hypothesize that the dismissal of employees due to partial work disability in positions

within the public sector is very unlikely. Even in the private sector, such actions are both ethically and sometimes legislatively unacceptable.

Overall, the psychological and behavioral parameters of the workplace and of the social structure of the work group create an effect on a variety of factors ranging from health to work motivation and efficiency of labor. Their role is often analyzed in combination with the previously described more general physical characteristics of the work relationship through the concept of employee well-being. An overview of the literature related to this concept is presented in Savin (2009), where its role is linked to the rehabilitation and retention of employees.

Academic literature generally supports most of our hypothesized relationships between the psychological, physical and background factors and employee work health and work ability. Many of these relationships, such as the dependency of the work ability on the type of physiological illness can be included into a quantitative model, while other factors such as the qualitative workplace characteristics present a significant challenge and may inhibit the model from obtaining high predictive power. Within the next section we proceed to a more applied level of analysis where the model specification from Savin (2009) is reviewed, augmented using the findings of this section and finally a research plan is developed.

2.3 Disability pension system in Finland

Due to the fact that the data used in this paper includes the information from a part of the Finnish working population, it is important to understand the disability pension security system in Finland and also consider several types of disability pensions which will also be present in the data set under analysis. This section provides a brief overview of the Finnish disability pension system from a practical perspective.

Disability pensions in Finland guarantee a fair pension compensation for the employees who can not continue working in full due to loss of work capacity. These types of pensions are fairly common, due to the fact that both illness and accidents may often lead to loss of work ability through a variety of diagnoses. In the organizations insured by the Finnish State insurance institution up to 9% of pension spending is associated to disability pensions.

Disability pensions are available in several different categories, depending on the state of the work ability of the individual. Each category has slightly different rules and compensation size which apply to it. Below, the main pension types are presented

2.3.1 Full disability pension

The criterion for granting a full disability pension is the loss of at least 60% of the working capacity of an employee. The disability pension may be granted either permanently or for a fixed term with a rehabilitation plan to restore the working ability of the individual. Since the employee moves to disability pension earlier than at the required retirement age, his accrued pension on the basis of his work relationship is lower. For this reason the disability pension adds an additional projected pension component to the accrued pension, which estimates the amount of pension the employee would accumulate if he or she worked until the age of 63. This extra component is only included if the employee has earned sufficient income for the past 10 years of employment. As a result, the employee receives a fair equivalent of the pension he or she would have obtained during his full working life. Once the employee reaches the age of 63, the disability pension is terminated and replaced by an equivalent old-age pension.

2.3.2 Partial disability pension

The criterion for granting a partial disability pension is the loss of at least 40% of the working capacity of an employee. The monetary size of the compensation in this case is half of the permanent disability pension, but may continue to work part-time as long as he or she does not receive more than 60% of full-time salary. Partial disability pension may also include a rehabilitation plan. Partial disability pension may also be gradually moved to full disability pension if the working ability of the individual decreases further. Once the employee reaches the age of 63, the partial disability pension is also terminated and replaced by an equivalent old-age pension.

2.3.3 Vocational rehabilitation

This benefit is granted if there is a threat to working capacity, which can be treated with pre-emptive methods. Vocational rehabilitation was introduced in 2004 and still remains a fairly rare category of disability pensions.

The decision on the granting of a disability pension is important, because it has both a long-term effect on the employee's life and also a financial effect on both the employer and the

employee. Due to this reason, the decision on the necessity of a disability pension is not made by a physician, but instead by the State Treasury, which may review the physician's statement and other supporting material including the health requirements of the position filled by the employee suffering from work disability. This guarantees the fact that employee receives fair treatment and the characteristics of the workplace are taken into consideration to distinguish differences in requirements (for example, physical) between different professions and positions within an organization.

Additionally there are several significantly rarer types of disability pensions and rehabilitation methods, however, they do not strongly affect the general picture and also the frequency of their occurrence in the data set under analysis will be so low that they will be excluded from the analysis. As a result, the abovementioned types of disability pensions will split the pension types into full versus partial and into rehabilitation allowance versus only disability pension. This set of pension types allows accommodating for a variety of health problems affecting work ability and offers employees an opportunity to maximize their work potential and support their health state.

2.4 Statistical model review and selection

The key aim of this thesis is the development of a statistical model to produce an indicator of disability pension risk, which could be applied by government offices and agencies in both financial forecasting and possibly in pre-emptive actions in high-risk groups. There is a variety of modeling techniques available and for this reason the initial selection of an appropriate modeling technique is necessary. In his previous research (Savin, 2009) the author has established a basic model structure and has proposed the use of a state-space model for describing the link between background variables, sickness absence data and disability pension risk. In this section, the model specification developed in the previous study will be taken as the basis for further development and augmentation. Initially, the basics of Markov chains and state-space models will be reviewed. In the second part of this section, a simpler model will be presented as a point for comparison. Finally, issues related to model testing and benchmarking will be discussed. As a result, this section of the literature creates a solid basis for the understanding of the statistical methods and approaches, which will be used in this thesis.

2.4.1 State-space models and Markov chains

State space models describe stochastic processes, where the model can be defined as a set of states $S = \{s_1, s_2, ..., s_n\}$ and any individual model state can always be described by a state in the state space *S*. The dynamics of the model are then defined by adding transition probabilities between each possible combination of states. According to previous research (Savin, 2009) this type of a model provides a sufficiently simple, but powerful framework for the analysis of individuals and populations especially from the perspective of health and work ability characteristics, because individuals can be easily classified into suitable states according to their employment relationship and health state (e.g. healthy, temporary sickness, temporary disability pension, permanent disability pension, death can be used as a set of states for a given individual). The model can be further simplified by removing any variation in transition probabilities, which creates a time-homogenous Markov chain process – one of the most basic state space models. This was selected as the modeling method in Savin (2009), because it provided opportunities for aggregation of individual level models.

Markov chains are stochastic processes, which are characterized by a set of states and transition probabilities between them. The key characteristic of a Markov chain is the fact that the transition probabilities from the current state are independent on the past states of the process (Markov property). In other words, the history of the system is irrelevant when determining the transition probabilities. Mathematically, this relationship can be summarized as the transition probabilities conditional on the current state being equal to those conditional on the full history of the system (Vanden-Eijnden, 2009):

$$P(D(n+1) = d \mid D(n) = d_n) = P(D(n+1) = d \mid D(n) = d_n,\ D(n-1) = d_{n-1} \ldots D(0) = d_0), \qquad (2.1)$$

where *D(n)* is the state distribution at time *n* and *P(D(n+1)=d)* is the probability to obtain state distribution *d* at time *n+1*.

The transition probabilities, on the other hand, can also depend on variables present in the model and can be described by transition function. Original analysis in Savin (2009) was limited to time-homogenous Markov chains with constant transition probabilities, where the model was described by a transition matrix (Grinstead and Snell, 1997):

$$P = \begin{pmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{pmatrix}, \tag{2.2}$$

where p_{ij} is the transition probability from state i to state j.

The selection of a time-homogenous state-space model in Savin (2009) set a strong limitation on the types of variables which could be included in the model. Inclusion of variables into the model was performed through the inclusion of new states, which artificially separated individuals according to a background variable into separate states. For example, a basic model from Savin (2009) is presented on **Figure 5**.

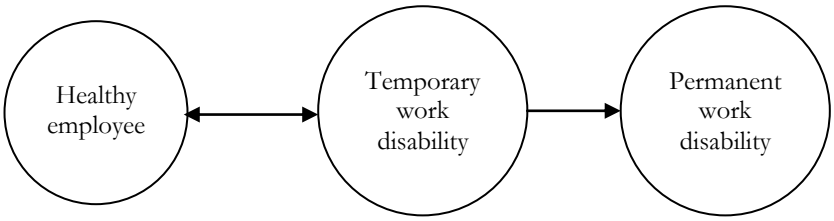

Figure 5. Basic Markov chain model (Savin, 2009)

Savin (2009) specifies a rigid process for the inclusion of new variables into the model. Firstly, the variable has to be categorical and should not have a very high number of possible values. Secondly, transitions between different states of the categorical variable must be considered (while it may be relatively easy to model the transition probabilities between different values of gender, it may be much harder to model similar transitional probabilities between different weight groups or other variables which can change quite often). As a result, each variable adds several new states to the model. A set of new variables may create a very large number of states as each feasible combination of variable values must be mapped as a new model state. **Figure 6** illustrates the inclusion of sickness absence variable into the model.

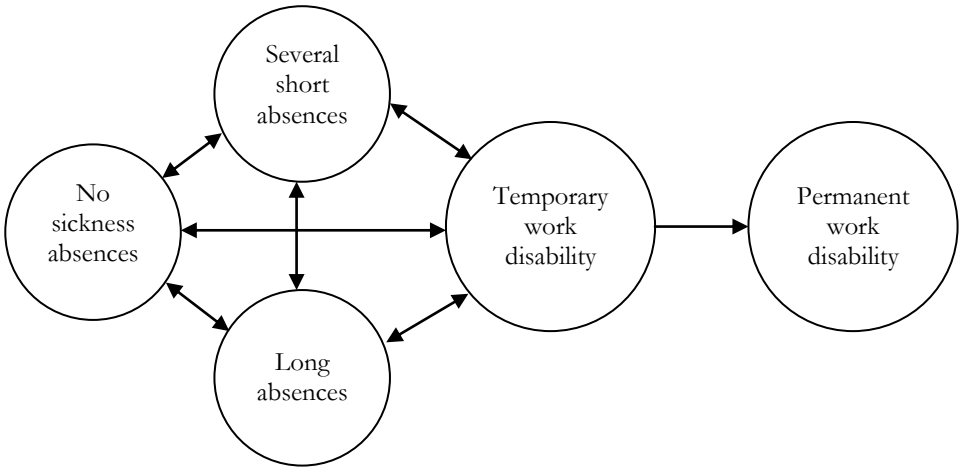

Figure 6. Final Markov chain model (Savin, 2009)

This limitation in Savin (2009) was justified by the possibilities of application of the model to aggregate level data, since aggregation of individual Markov chains is relatively simple. On the other hand, within the scope of this thesis, an individual level model will be the key focus and it will be used to determine disability pension risk levels for groups of individuals. In this way, it can be argued that the role of variables describing the background information about individuals will be quite high and additionally the requirement of the applicability of the method for analysis of aggregate level data is no longer present. Due to these reasons, within the scope of this study we will allow the transition matrix to vary depending on values of other variables. As a result we will use a more general transition matrix, which will include a set of transition functions, determining the transition probabilities for specific individuals:

$$P = \begin{pmatrix} p_{11}(X) & \cdots & p_{1n}(X) \\ \vdots & \ddots & \vdots \\ p_{n1}(X) & \cdots & p_{nn}(X) \end{pmatrix}, \tag{2.3}$$

where $p_{ij}(X)$ is a transition function between states i and j, which is dependent on a set of variables X.

Determining the nature of these transition functions and the variables included in them will be one of the main challenges in the modeling process.

2.4.2 Alternative models used

The modeling of disability pensions using state space model is an attractive solution, but it also requires a fairly good understanding of the process under analysis in order to successfully

specify the states of the system (in this case, the individual's health state). Poorly chosen state specifications can be problematic and for this reason a simpler model will be reviewed in this section.

Disability pensions are events, which can be characterized by a certain probability. This probability clearly depends on a variety of background factors and also may be linked to the sickness absence numbers for the individual. As a result, it is logical to hypothesize a direct statistical relationship between the probability of a disability pension event and a combination of background variables and sickness absence data. Such direct relationships can be suitably described by regression models. Due to the fact that we are dealing with probability estimation a logistic regression model is one of the more suitable models to link these variables (Hosmer and Lemeshow, 2000). In a logistic regression model, the variables are fir to a logistic curve function. Since the produced values never exceed 1 or fall below 0, this model is ideal for modeling probabilities of observed events. As a result, in a logistic regression model, the following setup is used to estimate event probabilities.

$$p(x_1, x_2, ..., x_n) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...\beta_n x_n}}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+...\beta_n x_n)}}, \tag{2.4}$$

where $p(x_1, x_2, ..., x_n)$ is the probability of disability pension given variable set $x_1, x_2, ..., x_n$, β_o is an intercept coefficient and β_i is the regression coefficient for variable x_i. The parameters of such model can be conveniently estimated using the maximum likelihood method.

Having a model for disability pension probabilities forecasting can be performed by estimating the probabilities of disability pension for each individual in the population. Depending on the probability a forecast is either set to be a positive disability pension event or lack thereof. To separate these events a threshold value is set for the disability pension probability. Usually this value will be at 0,5, which means that if the disability pension probability estimated by the model is above 0,5 then the model output is a positive disability event prediction. On the other hand, for the model can be more sensitive if this threshold value is decreased. However, decreasing the threshold value results in a significantly larger amount of type I errors, where a disability pension is predicted, but does not occur.

As a result, the logistic regression model can be estimated with very little effort and provides a simple alternative to the state space model. For this reason, this model will be applied in this paper, but will not receive the central role.

2.4.3 Model testing and benchmarking

In any practical application, a model must be judged by its forecasting power and performance. The fact that the model development process is sound and that the theoretical observations and frameworks are included in the model development process is insufficient to guarantee actual performance of the model on real data. Not only the model could turn out to be fairly weak, but also the hypothesized relationship between disability pensions and sickness absences could turn out to be insufficiently strong on an individual level. For this reason, model benchmarking is an important area of analysis within the scope of this thesis.

The strength of the relationship between the sickness absence data, background variables and the disability pension events can be analyzed by applying a regression model. This is where the previously discussed logistic regression model comes into play. It will be used to show the results which can be obtained using purely the direct relationship between sickness absences and disability pensions. After this, the state space model will be analyzed to determine if it is able to meet and possibly surpass the logistic regression model's result. If the state space model will surpass the logistic regression model, then the selected states are beneficial and realistic. On the other hand, if the introduction of states into the system description creates inferior results – the state selection will have to be revised.

Finally, the main benchmark criteria used are set from the practical perspective. The number of type I and type II errors on a validation sample will be estimated. Type I errors in this situation will be the incorrectly marked disability pension candidates and type II errors will be the non-identified employees with disability pensions.

The analysis of both type I and type II errors with a comparison to a benchmark model will allow to determine an objective performance measure for the state space model which will be developed. Additionally, qualitative analysis of the model capabilities will be used to comment on the possibilities for model interpretation and applicability of the model in actual decision making.

Section summary

The volume of literature in the related subjects has increased, but the discussed area still lacks research.

Some studies on aggregate level exist, but do not proceed beyond simple description of the relationship.

Psychological and behavioral studies indicate the role of background variables in the retirement and absenteeism decisions.

State space model is selected as the primary model and a logistic regression model will be used as a benchmark.

3 Descriptive Data Analysis

Before a full model can be developed, the data set used in this thesis was first analyzed. The reason for this was the previously mentioned complexity of appropriate state space specification. In order to create constructive states for the individual's health state, the internal data structure must be analyzed and the main patterns must be considered.

In this MSc thesis project, an individual level data set was obtained for analysis and for the development of specifications for the model. This data set was comprised of the information related to disability pensions in the population of the employees of the Finnish government offices and agencies, which was provided by the Finnish State Treasury and of information related to monthly sickness absences in years 2006-2008, which was provided by the Finnish Ministry of Finance. The data set comprised of a large number of individuals and observations and included certain specific issues, which had to be tackled. In this section, the descriptive analysis of the data set will be presented and the key challenges related to its structure will be identified. The structural changes made to the data set and their rationale will also be discussed.

3.1 Variable description and initial modifications

The data was available in a spreadsheet format, which included personal identification numbers to identify individuals on rows. The variables were specified on a set of columns. Some of the more important column groups were:

- *Gender and year of birth*
- *Agency code* – represents the public body at which the individual is employed
- *Employment relationship code* – represents a specific employment relationship. An individual can belong to more than 1 position in an organization during the 3 years period and thus there can be more than 1 data line per individual.
- *Sickness absence type* – this variable has a value of 019 for absences where the salary is paid and 082 if it is not. This can also lead to more than 1 data line per employee.
- *Number of sickness absence days and periods in each month between 2006 and 2008* – this variable was in a cumulative form, where each month included all absences from the beginning of the year to the current month.

- *Disability pension type* – if the individual was granted disability pension this shows if it was permanent or temporary. Also this may show if rehabilitation measures were taken or not.
- *Disability diagnosis* – shows the key type of sickness due to which the disability pension was granted. The general types are mental illness, circulatory illness, disability related to moving limbs and other disability reasons.
- *Starting date of the disability pension*
- *The monetary value of the disability pension*

More detailed information on the variables and their symbolic representations can be found in the appendix.

The first observation, which was already clear from the variable explanations, is the fact that the data set includes multiple data lines per individual and these have to be merged. This was the first step, which was performed. The paid and non-paid disability pension types were placed into different columns (as a result each individual now had 2 time series of sickness absences – type 019 and 082, but they were both present on one data line).

The merging of several employment relationships per individual was a larger dilemma, because the process would result in data loss and possibly incorrect data merging. It is clear that if a person is employed in several positions, he could report his sickness absences in both, while on the other hand the individual may only register his sickness absences on one of them. After a preliminary analysis of the data it was found that the latter case, where the individual only registers absences on one of the employment relationships was more common. For this reason, the lines were merged in such a way that all sickness absence data related to a single individual was summed over all of his or her employment relationships.

Additionally, the sickness absence data had to be transformed from a cumulative to a frequency perspective. This simply meant that for each of the years the 1st month remained the same, while each next month's absence number would be the difference between the cumulative values. This was a simple procedure, but as it will be shown later, it revealed information about errors within the data.

3.2 Data errors and specifics

In order to analyze the integrity of the data set, several logical tests were performed. Firstly, a test for non-negativity of sickness absence numbers for each month was performed. It was fairly interesting to notice that this test already revealed the fact that a large number of sickness absence records was deficient. The reason for this, which was suggested by the Finnish State Treasury, was the fact that the data set was not developed in such a way that all possible data entry errors are checked on monthly level, but rather to show the annual-level statistics. As a result, the monthly cumulative figures could be distorted, which resulted in negative sickness absence results for some single months. Due to the abundance of the data and due to the fact that an assumption of the randomness of these errors could be made, the observations with negative values were removed from the data set. In case of such situation the whole individual was removed in order not to distort the data.

Additionally it was found that for some individuals the number of sickness absences exceeded the total number of days in a year. The quantity of such observations was extremely low and some of them were a result of the merger of several employment relationships, where individuals did report their sickness absences in all of the relationships.

As a result of the abovementioned changes, the final size of the data set was at 98’333 individuals. The initial size of the data set was 141’114 lines, 20’834 of which contained negative values and 21’948 of which were removed due to the fact that they were merged. The total number of disability pension events (of all types) was 2060 in the final data set. This corresponded to 2,1% of all of the individuals. It is important to note that the disability pensions did not all happen within the range of 2006-2008 and some of them lay quite far outside of this region, which explains a fairly large percentage of disability pensions.

Finally, it is also important to note that within the data set only individuals with at least 1 sickness absence or a disability pension were recorded. This means that if an individual did not have any sickness absences during the 3 years, he would not be included into the data. This could be understood as a limitation, but also as a positive sign, since, from a subjective viewpoint of the author it is very unlikely that a person does not fall sick at least once during a period of 3 years. This means that the individuals who have no sickness absence records most likely simply do not report them within the organization. The exclusion of such

individuals is not a problem, since their data would not reveal any truthful information in any case.

3.3 General statistics

The analysis of the data started, in a very similar way to the literature analysis, with the separate overview of data related to sickness absences and different pension types in order to get better knowledge and understanding of the data at hand. This process and its results are outlined in this section.

3.3.1 Sickness absences

The sickness absence data was represented by the absence type and by 12 monthly figures for each of the 3 years. On average each employee was absent due to sickness for 0,723 days per month with a paid sickness leave and only 0,00978 days per month with unpaid. This indicates the rarity of sickness absences of type 082 (unpaid).

The distribution of occurrences of different sickness absence types was also analyzed during each year. **Figures 7, 8** and **9** show the distributions of individuals with a given number of short sickness absences during each of the 3 years. Logarithmic scale is used due to a very large difference in the quantity of large and low absence numbers in each of the years (e.g. 30'000-40'000 have no absences).

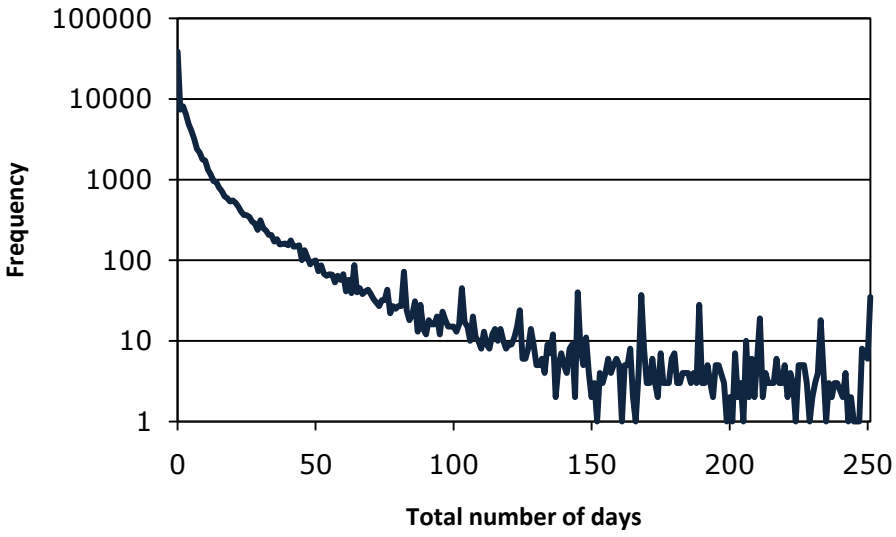

Figure 7. Paid sickness absences in 2006 (frequency of observations)

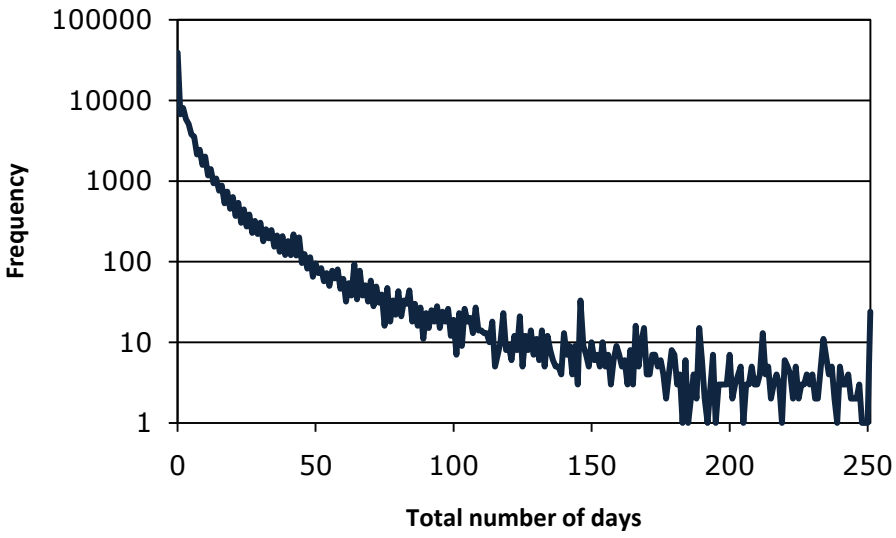

Figure 8. Paid sickness absences in 2007

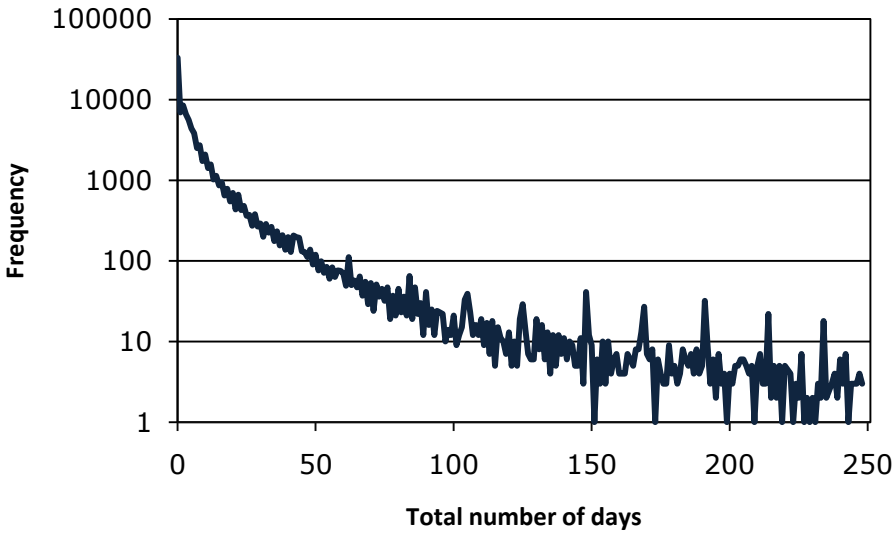

Figure 9. Paid sickness absences in 2008

There are a large number of individuals, who take only a couple of days or no days at all of sickness leave during a year and as the number of days grows, the number of individuals decreases exponentially. Nevertheless, a fairly interesting observation can be made. There are large spikes, which are present approximately every 18-24 days and end at 250-252 days. This indicates individuals on long-term sickness leaves who are granted sickness leaves on a monthly basis and may even span the full 250-252 working days in a year.

The unpaid sickness leaves are, as it was previously stated, very rare. The percentage of individuals taking this type of sickness leave does not exceed 0,25% in any of the 3 years. As a result, their distribution consists of a majority of individuals having 0 unpaid absences and

very few individuals having other values. An example of the distribution is presented on **Figure 10**.

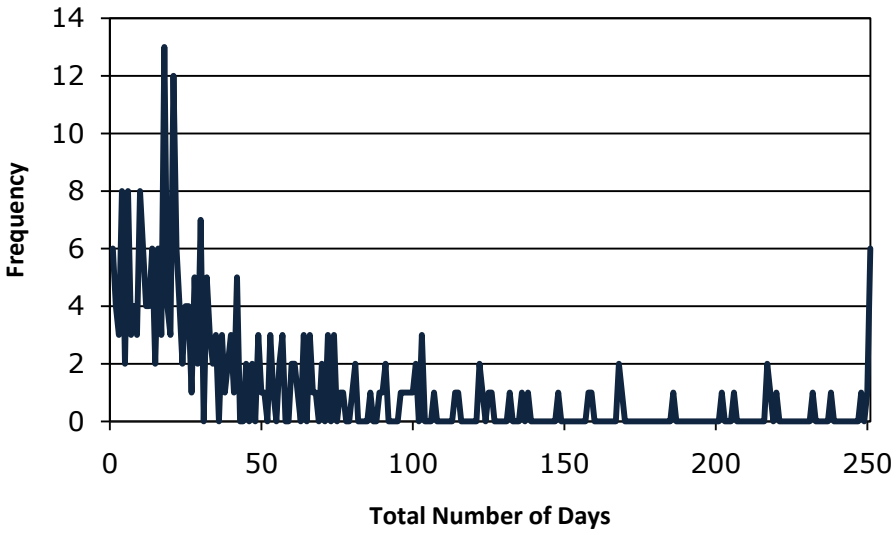

Figure 10. Unpaid sickness absences in 2006 (observation for 0 days is omitted)

Between years there are no significant variations in sickness absence data, but since our model will operate with these values on monthly levels, it is very important to consider the cyclical variations in sickness absences within a year. The variations in sickness absences could both lead to the procedures related to recording of these absences, but also to the annual cycles in certain mild illnesses such as influenza. Monthly numbers of sickness absence days per individual were recorded and averaged over the 3 years of data. As a result, **Figure 11** illustrates the results which were obtained for paid sickness absences.

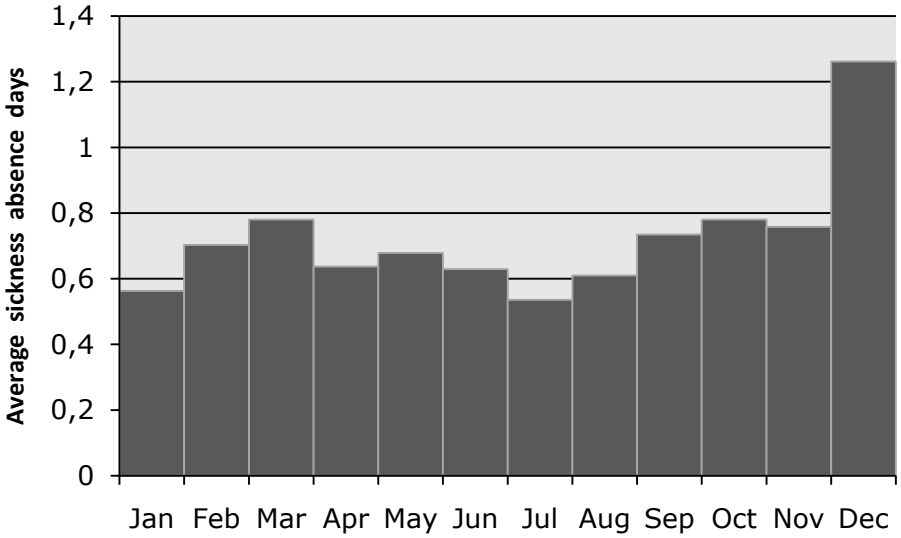

Figure 11. Average number of paid sickness absences per month in days

It can be clearly seen that there are large variations in the sickness absences throughout the year. Firstly, the beginning of the winter in December is signified by a large spike. This can be associated not only with the changes in weather, but also with the fact that many sickness absences may be related to other forms of absenteeism resulting from a large number of holidays in the month. During the summer, on the other hand, due to the reason that a large number of employees are on holidays and due to positive weather conditions, the situation is very different. The number of sickness absences registered in July is less that 50% of those in December. This leads us to a conclusion that in the modeling process it may make sense to use sickness absence numbers relative to the population average for individuals instead of absolute ones to account for the annual cycle.

In the case of non-paid sickness absences the situation is very different and it is illustrated on **Figure 12**. These absences are taken much more often during the summer months and the general variation in their number is lower. Here, the reason could be the voluntary nature of these absences. Since they are not paid for, the employee may not need any confirmation from medical staff and could use this opportunity as a non-paid vacation. Naturally, this is only a hypothesized argument.

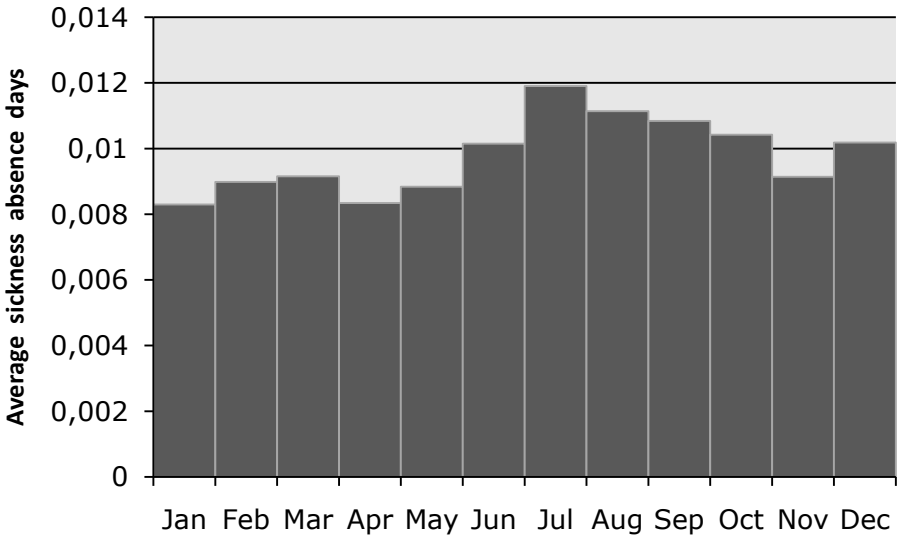

Figure 12. Average number of unpaid sickness absences per month in days

On the basis of the basic analysis of sickness absence data we can say that the type 019 paid absences are much more frequent and are most likely more related to common causes of physical illness. There are both short-term and long-term absences in this category, which means that the paid sickness absences have to be modeled not only using their total quantities,

but also using their average durations (the average duration of a paid sickness absence was 4,86 days).

The non-paid sickness absences are much rarer and the causes for their variation can not be fully objectively described. Their average duration is also longer – 10,49 days. These sickness absences may relate to specific illnesses or motivational problems, but the relationship with disability pensions must be further analyzed.

3.3.2 Disability pensions

The number of disability pensions within the sample was fairly high and it constituted 2,1% of the total sample population. In order to better understand the main causes and structure of disability pensions in the population, an analysis of several parameters of the disability pensions is performed.

The first parameter of core interest is the age. It is clear that the age will have a strong effect on the transfer probabilities within our model and it was hypothesized in the literature review that higher age values would result in significantly higher probabilities of disability pension. This hypothesis is partially supported by the data. **Figure 13** shows the age distribution within the set of 2060 disability pensions present in the population under analysis.

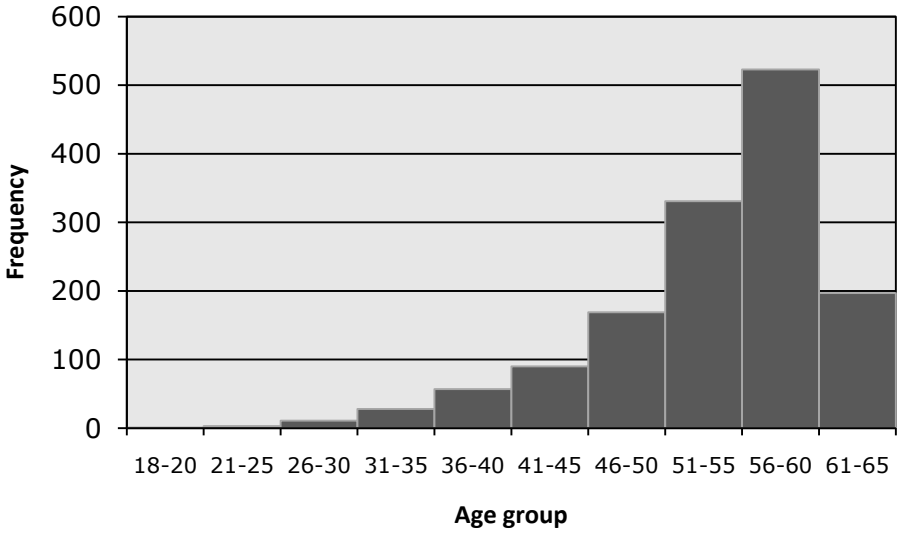

Figure 13. Distribution of age at pension start

It is quite clear that the disability pension probability increases up to the age of 56-60. The likelihood of serious illness increases with age. However, another effect in the population is the fact that amount of disability pensions significantly drops for the population over 60 years

old. The reason could be the alternatives available to the employees at this age. After all, the employees may have an option to simply take normal early pension based on their age. The process may be significantly simpler and as a result may be more desirable from the perspective of the individual. On the other hand, some may perceive the financial benefits of disability pension to be significant enough to justify the more complex procedures and still are granted disability pension at age above 60 (this is the 3rd highest age group on the graph, after all).

The disability pension granting process is not short and this affects the distribution of disability pensions granted within each year. Holiday seasons will most likely reduce both the number of individuals applying for disability pensions and the processing capability of the issuing organization. This means that the fluctuations within each year will also be significant. The resulting seasonal pattern can be seen on **Figure 14**.

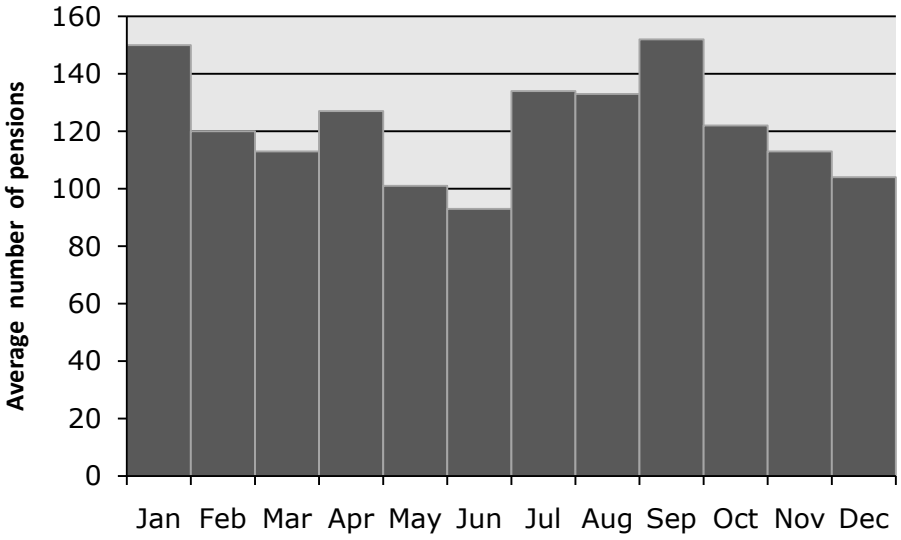

Figure 14. Average number of disability pensions per month

It is quite clear that the late spring and first summer months have a significantly lower number of disability pensions being granted and they are followed by a spike in July, August and September. A similar slump can be seen in December, which is followed by a spike in January. Similarly to the sickness absences, there is nothing surprising about this seasonality pattern, but it has to be considered later in the model development process.

The sickness absence data and disability pension data were gathered from two separate sources. The sickness absence data covered a period from 2006 to 2008, while the disability pension data covered a significantly larger period. For this reason only the suitable pensions

had to be kept within the data set and the older and later parts of the data had to be filtered. By removing all the pensions outside the 2006-2008 region, the number of disability pensions in the sample decreased from 2060 to 1464. The density of observations by month and the filtered tails are illustrated on **Figure 15**.

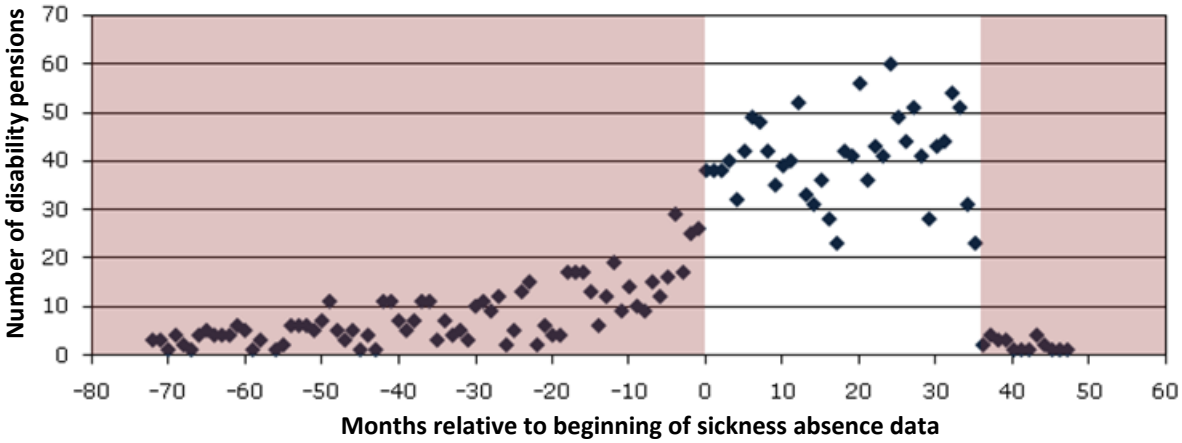

Figure 15. Pension starting point relative to the sickness absence data (frequency vs. starting point)

In the interpretation of the graph we have to be careful, because the data set includes only individuals who were working and had sickness absences between 2006 and 2008. For this reason, the highest density of disability pensions is exactly at that time. The later disability pensions could be infrequent due to the fact that this period would be very recent and the statistics have not yet been fully reported. The disability pensions present in the previous periods are most likely non-permanent disability pensions, where the individuals managed to recover and returned to work during the period of 2006-2008.

The disability pension data also identifies the type of the disability pension and the type of diagnosis the employee had. This data could be valuable in the modeling process of different disability pension types. There are 4 main groups of pensions present in the data (described previously in the analysis of the Finnish disability pension system):

- Full disability pension with rehabilitation support (type 8)
- Disability pension with partial rehabilitation support (type 9)
- Permanent full disability pension (type S)
- Partial permanent disability pension (type Z)
- Personal early retirement scheme (type Y)

Additionally, the diagnoses of the employees are grouped into the following categories:

- Mental illness (1MT)
- Illness of the circulatory system (2VK)
- Illness or disability related to moving limbs (3TU)
- Other diagnoses (4MU)

There is also a more thorough classification of these diagnoses; however the quantity of observations for each of the subclasses is much too low for further analysis. The distribution of disability pension between different pension types and diagnoses is presented in a tabular form in **Table 3**.

Table 3. Pension types and diagnoses

	Diagnosis				
Pension Type	**1MT**	**2VK**	**3TU**	**4MU**	
8	73	4	45	38	**160**
9	302	23	166	172	**663**
S	210	80	143	205	**638**
Z	130	35	241	190	**596**
Y	1	0	1	1	**3**
	716	**142**	**596**	**606**	

Several observations about the distribution of different pension types and their relationships with diagnoses can be made. Firstly, partial and full rehabilitation schemes are offered more often to the employees suffering from mental illness, which is natural, because a large volume of mental sickness problems can be alleviated through rehabilitation (Burton, Schulz, Chen and Edington, 2008). On the other hand, the dominant pension type for employees with injury related to limbs is permanent disability pension, which is also logical. What is interesting is also the fact that generally, the largest diagnosis group is mental illness, which further emphasizes the opportunities for preemptive treatment of employees. Nevertheless, even for other types of diagnoses there is a mediocre probability of receiving some type of partial or full rehabilitation support within the disability pension. Finally, the personal early retirement scheme is such a rare type of pension that it will not be considered in further analysis.

Having analyzed the parameters of the sickness absence variables and the disability pension data in separate it is now possible to consider the again insights into their relationship. This will be achieved through an event study of disability pensions in the next section.

3.4 Event study of disability pensions

In order to obtain better insights into the possible parameters which will go into the further model and to formulate possible states for the state space model, an event study related to disability pensions within the given population was performed. Within this section the additional data set modifications will be discussed first, after that analysis of the disability pension event will be performed. Finally, two different types of processes leading to disability pensions will be identified in the data and a hypothesized explanation for this division will be presented.

3.4.1 Data modification

The first challenge related to the event analysis was again the data structure. Since the time series related to sickness absences were recorded in absolute time and the sickness absences took place at different moments, the time series for each individual with a disability pension had to be reorganized. The desired organization was such that the time parameter of the series would be relative to the pension starting time. For example, a time parameter of 1 would contain variables related to 1 month prior to a disability pension. This transformation was not difficult to perform, but it created further limitations within the data set.

Since the disability pensions were fairly evenly scattered within the period between 2006 and 2008, this meant that the history available for each disability pension observation was not the same. Observations of disability pensions, which were located in the early part of the data (e.g. 01/2006) did not have very much pre-pension history, while others could have up to 3 years. On the basis of the modeling horizon, it was decided that the minimum number of history required for an acceptable disability pension observation would be 2 months and history up to 12 months will be analyzed. All of the observations with a shorter history were filtered. As a result, 1388 observations remained. On average, each disability pension observation contained 10,4 months of history, which is very close to the ideal requirement of 12 months.

3.4.2 General results

The key part of the event analysis was the observation of sickness absence patterns prior to a disability pension event. The key role was initially given to the paid sickness absences, since the number of unpaid sickness absences was very low. **Figure 16** below illustrates the average number of paid sickness absence days per month prior to a disability pension.

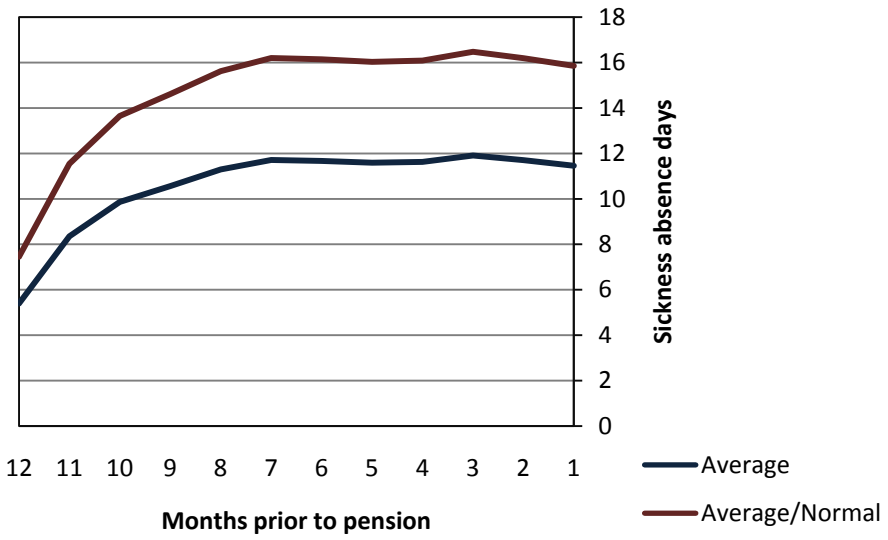

Figure 16. Average number of paid sickness absences days per month

The initial observation here is very clear. The blue line shows the average number of sickness absences per month, while the red line indicates the ratio between this figure for an average individual and for an individual with a disability pension coming up. So, for example, 1 month before the disability pension event individuals have around 12 sickness absence days per month, which is 16 times larger than normal. The number of sickness absences prior to a disability pension on average clearly follows a gradually increasing and saturating pattern. The saturation level is around 12 sickness absences per month, which is 16 times the figure for an average individual. This level is also quite surprisingly reached as early as 7 months prior to the actual disability pension, while even 12 months prior to the pension, the sickness absence level is 7-8 times the normal level. The behavior of the number of distinct sickness absence periods and their length is presented in **Figures 17** and **18**.

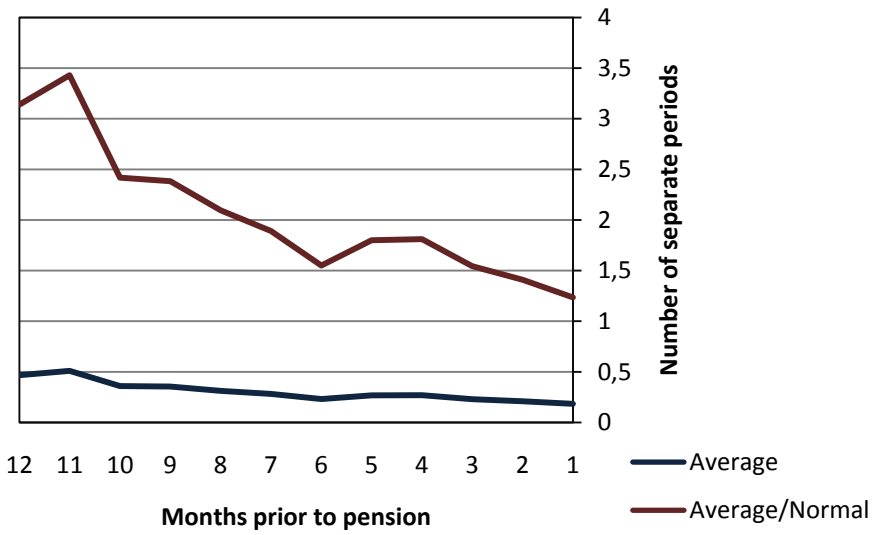

Figure 17. Average number of paid sickness periods per month

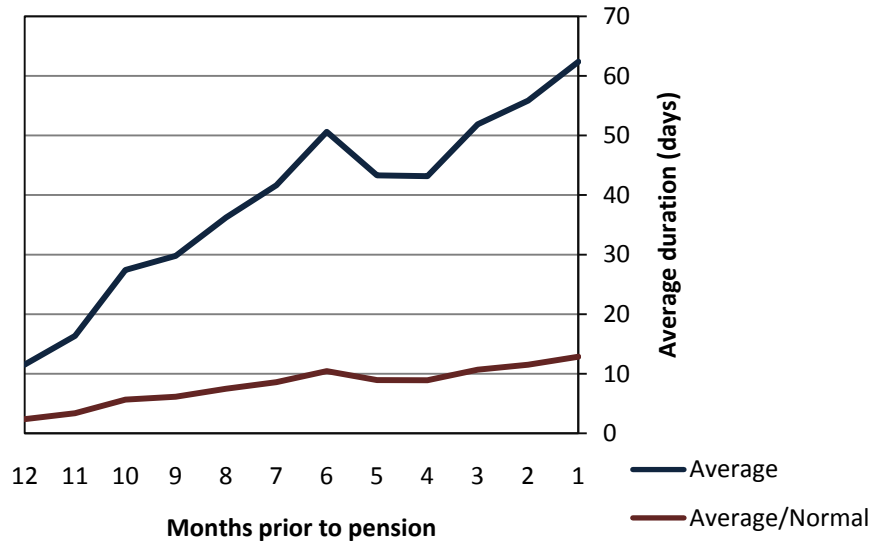

Figure 18. Average duration of sickness absences

The relationships here are again very clear. Despite the fact that the number of sickness absence days is very high and increasing, the number of distinct periods is decreasing and reaches almost the normal level just prior to the disability pension. This means that the average duration of a sickness absence is dramatically increased. This can be seen on the last graph, where the sickness absence duration reaches as much as 2 months a month prior to the disability pension event. This is over 10 times the normal level.

The observations related to paid sickness absences are very promising and indicate a very clear pattern of increasing quantity of sickness absence days and increased duration of

sickness absence events during the 12 months prior to the disability pension event. Given the fact that the model under development is positioned as a short- and mid-term model these results are very promising and may allow achieving significant forecasting power even with relatively simple model specifications.

Despite the low quantity of observations, the unpaid sickness absences were also analyzed in a similar way to the paid absences. The 3 graphs (**Figures 19**, **20** and **21**) corresponding to sickness absence days, periods and durations are presented below.

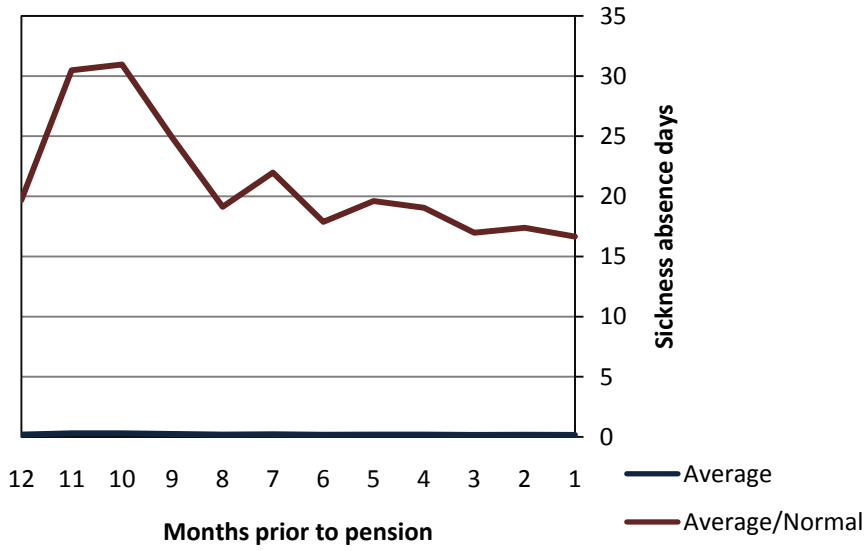

Figure 19. Average number of unpaid sickness absence days per month

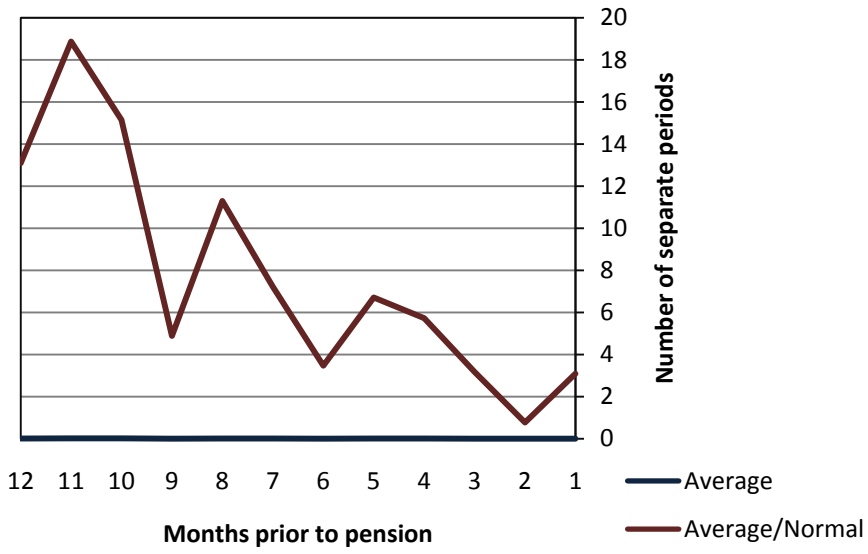

Figure 20. Average number of unpaid sickness periods per month

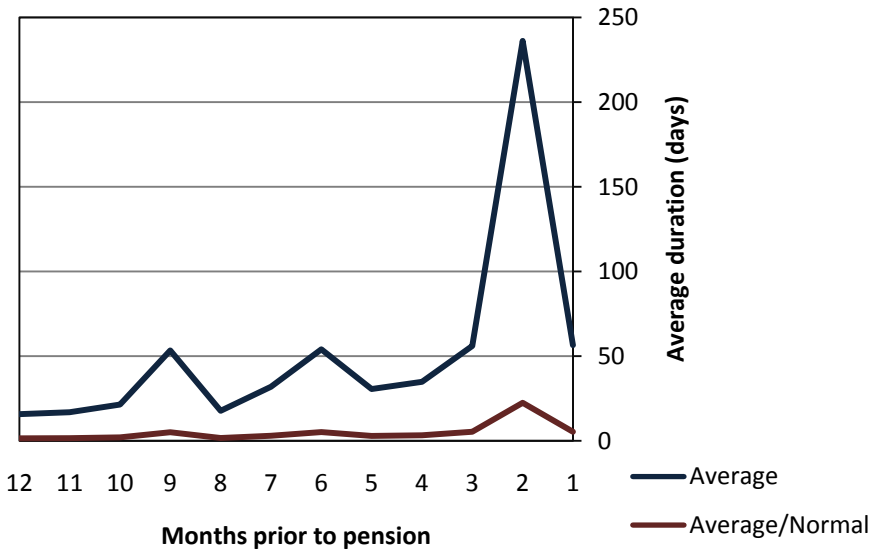

Figure 21. Average duration of unpaid sickness periods

There are only several significant differences between the unpaid and paid sickness absence patterns. Firstly, there are more fluctuations in the patters for unpaid sickness absences, which are a result of a significantly lower number of observations in the sample. Secondly, the first graph shows that the increased number of unpaid sickness absences is reached already before the 12 month threshold specified in the analysis. The threshold could be extended, but this would even further decrease the number of observations in the analysis due to the requirements for sufficient history. On the other hand, the trend with the decreasing number and increasing duration of sickness absences seems to be followed in a similar manner as in the paid sickness absence data.

3.4.3 Patterns within the population

The patterns in the aggregate data are fairly promising. However, for more precise analysis and model development internal patterns within different groups of the population have to be analyzed. The aim of this analysis is to reveal internal structure which would allow grouping of individuals into groups with possibly different sickness absence or other variable patterns. This would allow us to gain additional insight especially in the development of different employee states in the state space model proposed earlier in this paper. The focus in this section was placed again onto the analysis of mostly the paid sickness absences due to the high availability of observations for this type of sickness absences.

The first step in the analysis was to understand the distribution of different sickness absence frequencies in the 12 months of data prior to a disability pension event. The histogram for this distribution is presented below.

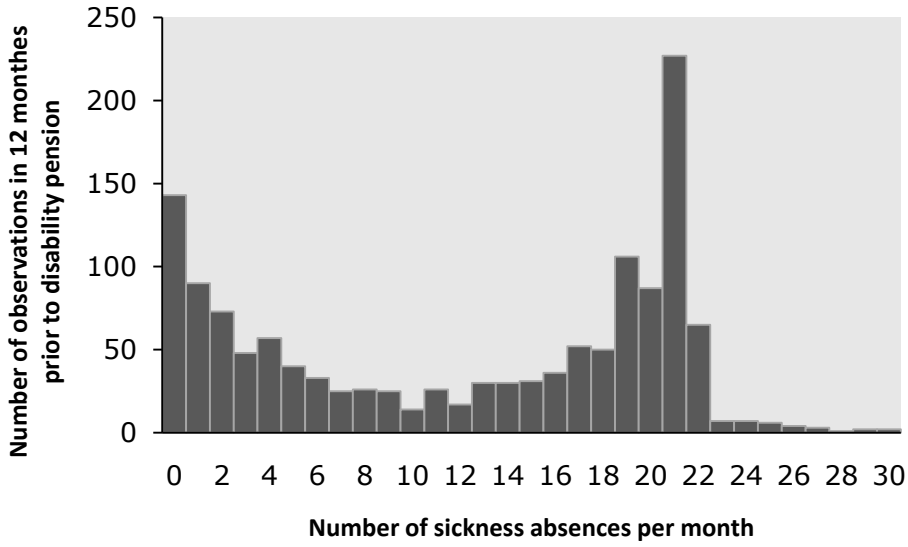

Figure 22. Frequency of sickness absence averages

The distribution in this situation is clearly bimodal with 1 peak being very close to 0 and another one at approximately 21 days, which is a full working month. This means that there are most likely two separate types of sickness absence patterns in the data. One is related to sudden (for example accident based, but this is not the only option) disability pensions, while the other is related to illnesses with gradual progression where the employee already experiences significant decrease in work ability over the 12 months prior to the disability pension.

Additionally it can be noticed that there are signs of normality in the two distributions (the left one, naturally, having only positive values may be a problem), which leads us to the hypothesis of a mixed bimodal Normal distribution of the sickness absence averages in 12 month. Fitting an appropriate distribution and comparing the likelihoods of different observations belonging to each of the components of the distribution will allow us to determine the cut-off value between the two types of onset of disability pension and analyze these two groups of observations separately.

First of all we would need to fit a suitable bimodal distribution to the sickness absence data. Due to the uncommon structure of the probability density function for bimodal distributions it

was decided that using a general maximum likelihood method would be most appropriate. As the basis for optimization the SAS nlmixed procedure was used.

The probability density function for a Normal distribution has the following form.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{3.1}$$

where μ is the distribution mean and σ is the distribution standard deviation.

A bimodal Normal distribution has the following probability density function, which is a sum of two Normal probability density functions.

$$f_{bimodal}\ (x) = pf_1(x) + (1-p)f_2(x) = p\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_1)^2}{2{\sigma_1}^2}} + (1-p)\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu_2)^2}{2{\sigma_2}^2}}, \tag{3.2}$$

where *p* is a scaling factor indicating the probability of an observation belonging to the first peak in the bimodal Normal distribution.

As a result, we can specify the likelihood function to have the following form.

$$\mathcal{L}(\mu_1, \mu_2, \sigma_1, \sigma_2, p | x_1, \dots, x_n) = \prod_{i=1}^{n} f_{bimodal}\ (x_i | \mu_1, \mu_2, \sigma_1, \sigma_2, p) \tag{3.3}$$

Finally a logarithm can be taken to obtain the log-likelihood function. The resulting function is then maximized with respect to the parameters of the bimodal Normal distribution to arrive at the optimal set of parameters for the two Normal distribution used in the mixed bimodal distribution.

The parameters of the resulting bimodal distribution are presented in **Table 4**.

Table 4. Bimodal Normal distribution parameter estimates

Parameter	Estimate
μ_1	0.794
σ_1	1.223
μ_2	19.345
σ_2	4.629
p	0.219

It can be clearly seen that the two normal distributions are fitted so that each one covers a peak on the original distribution. The two distributions in separate are presented in **Figure 23**.

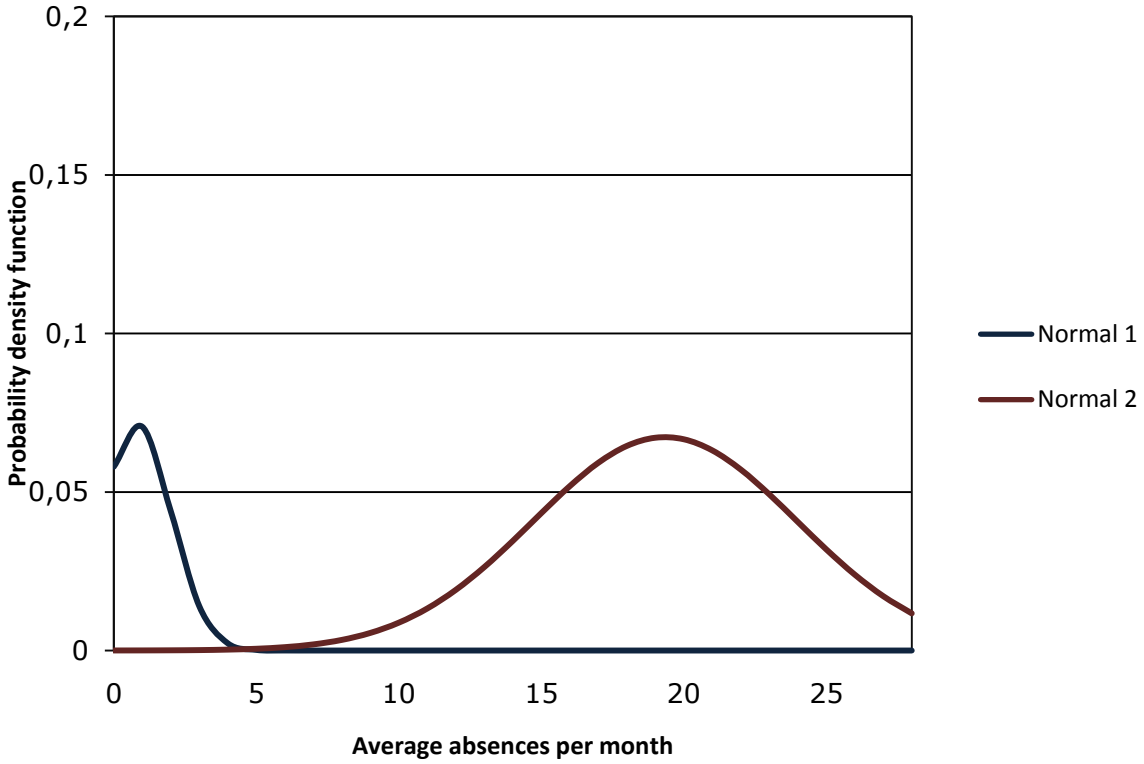

Figure 23. Two Normal distributions

The point where belonging to one distribution becomes more likely than belonging to the other can be used as the threshold value for classifying the observations into two groups. These distributions are then combined and plotted with the original distribution of data as it can be seen in **Figure 24**.

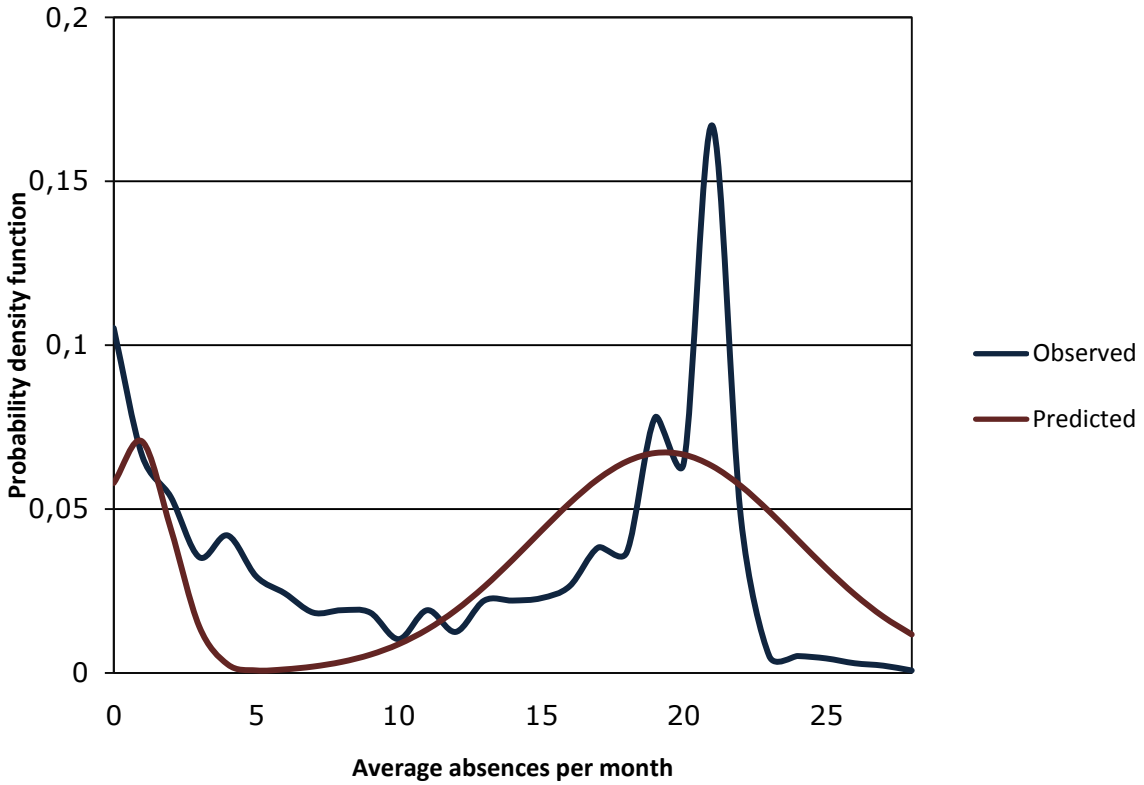

Figure 24. Bimodal Normal fit

There are several deficiencies in the fit which was created. Firstly, the left peak is placed in such a way that its standard deviation parameter is very low and this creates big errors around the value of 5. This makes it problematic to use this threshold value in further analysis, especially when an intuitive threshold point would seem between 10 and 15. Additionally, the second distribution has a high standard deviation and as a result does not fully match a large peak at the value of 21. This further contributes to the deficiency of the selected fit.

The first problem of a low standard deviation of the left peak is due to the fact that the first peak is very close to 0. In this way, since our distribution ends at 0 and the Normal distribution does not, the errors resulting from negative values are so large that in optimization they are minimized by minimizing the standard deviation of the first peak. Thus the Normal distribution fit is not necessary the optimal choice in this situation. On the other hand, the second peak is more or less normal and can be estimated with a Normal distribution. To resolve this problem, the Gamma distribution was used to describe the first peak and a Normal distribution was used for the second one. The Gamma distribution is a conjugate of the Poisson distribution. The resulting probability density for the bimodal distribution based on Gamma and Normal distributions is presented below.

$$f_{bimodal}\,(x) = p\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}x^{k-1} + (1-p)\frac{e^{\frac{-x}{\theta}}}{\Gamma(k)\theta^k}\,, \tag{3.4}$$

where μ, σ are mean and standard deviation of the normal distribution, θ and k are parameters of the Gamma distribution, $\Gamma(k)$ is a complete gamma function and p represents the probability of an observation belonging to the Normal distribution.

The likelihood function is then formed in the same way as previously. Using the SAS nlmixed procedure this likelihood function was maximized on the set of 1410 observations. The results are presented below.

Table 5. Maximum likelihood estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|
μ	19.365	0.107	1410	181.69	<.0001
σ	1.922	0.107	1410	17.89	<.0001
θ	20.234	1.530	1410	13.23	<.0001
k	0.325	0.013	1410	24.40	<.0001
p	0.433	0.016	1410	26.85	<.0001

The two distributions with the optimized parameters are presented in **Figure 25**.

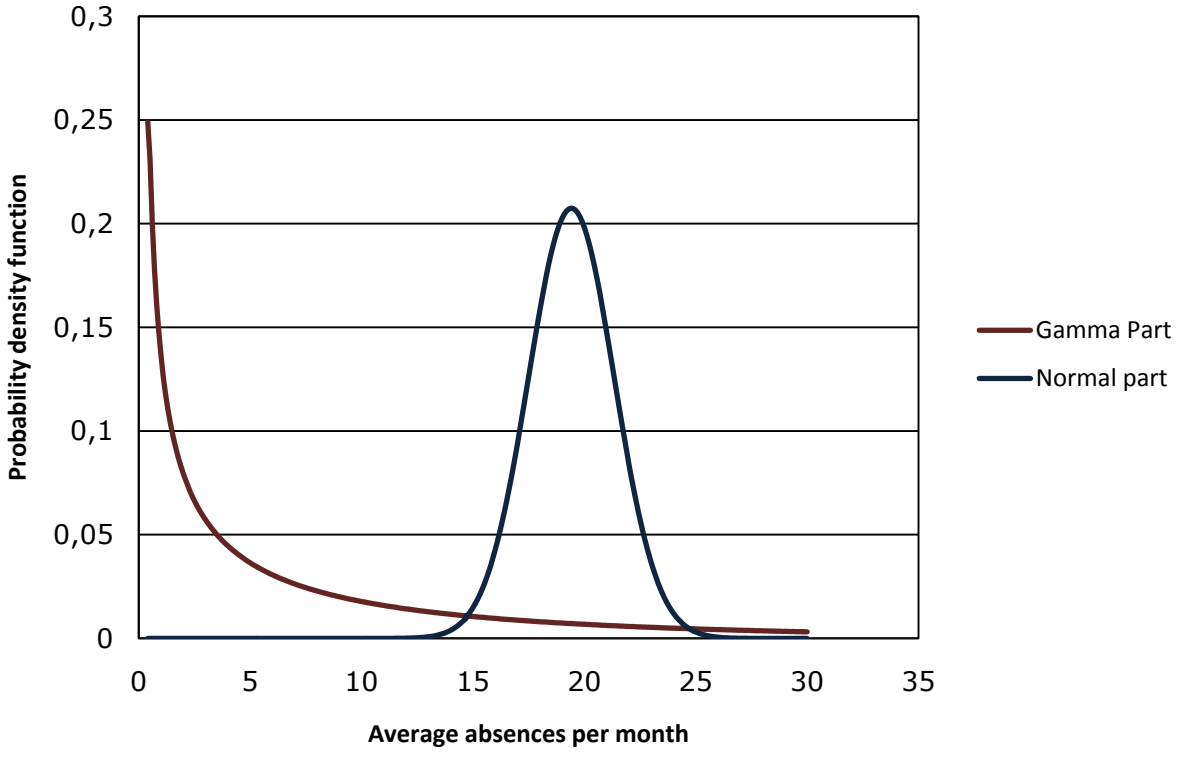

Figure 25. Normal and Gamma distributions

Finally, the mixed distribution is compared to the original distribution within the data set in **Figure 26**.

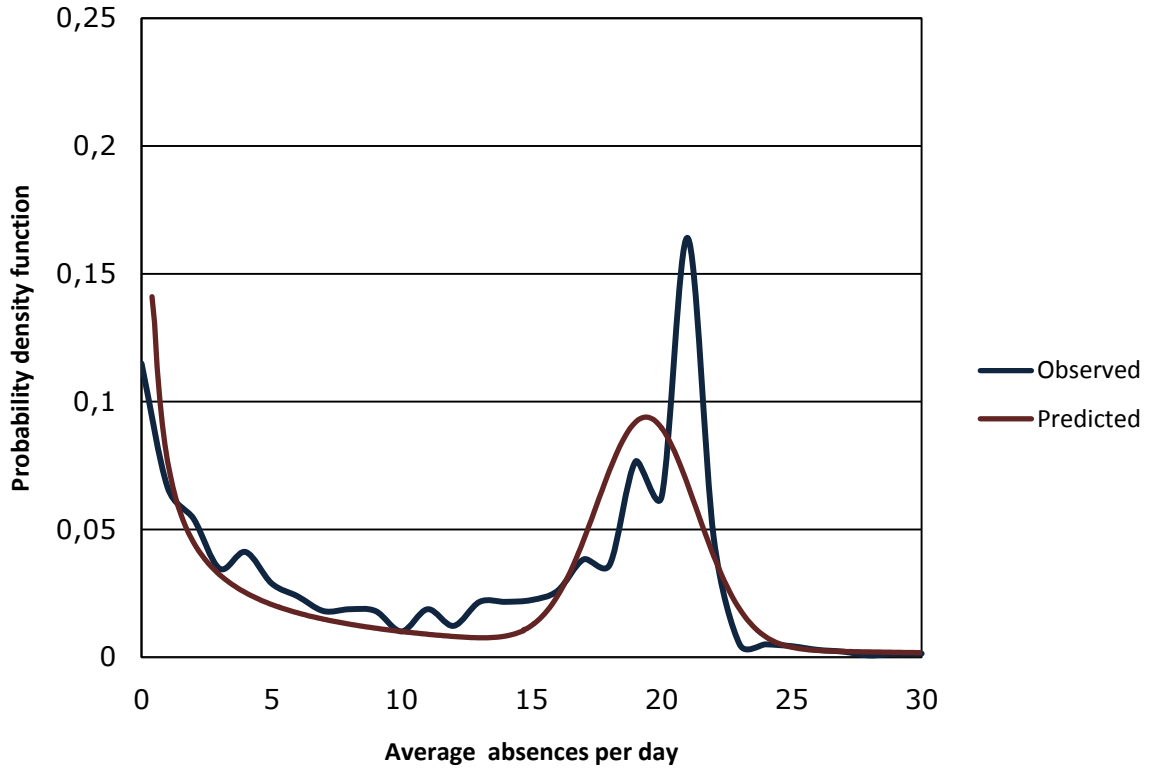

Figure 26. Mixed Normal and Gamma fit

We can see that the predicted distribution does indeed mimic the observed values quite closely. The size of the peak of the normal distribution is lower, but the overall shape is still quite good. Having fitted the bimodal distribution, we can now find a threshold value below which the likelihood of an observation belonging to the Gamma part of the distribution is higher than the probability of belonging to the Normal part. Above this value the likelihoods will be reversed. This value was estimated to be around 14,691 sickness absence days per month on average over the 12 month period prior to the disability pension. This value can now act as the point which will allow us to separate the disability pension events into two groups. The first group will be named "sudden", due to the fact that it represents the employees who had very little sickness absences prior to the disability pension. The second group will be names "progressive", because there is a clear sickness absence history prior to the disability pension. This will allow us to analyze the two groups in separate, which will be considered in the next section.

3.5 Analysis of progressive and sudden sickness absence patterns

Having established two different sickness absence distributions prior to a disability pension, we will proceed to analyze the sickness absence patterns within each of these categories. The procedure for the initial analysis will follow a very similar pattern to the General results part of this section.

3.5.1 Sickness absence patterns

Firstly, it is important to note that the two classes are very evenly distributed in the population. The progressive type of sickness absence distribution occurs 50,07% of the times, while the sudden occurs 49,93% of the times, which slightly deviates from the 43,3% value of *p* in the hypothesized bimodal distribution.

The next step in the analysis is the development of 12 month charts for average number of sickness absence days per month, average number of sickness absence periods and the resulting average sickness absence duration. Firstly, the graphs for progressive class of disability pensions are presented in **Figures 27**, **28** and **29**. We focus on the paid sickness absences, because once again they form the more significant part of the data set.

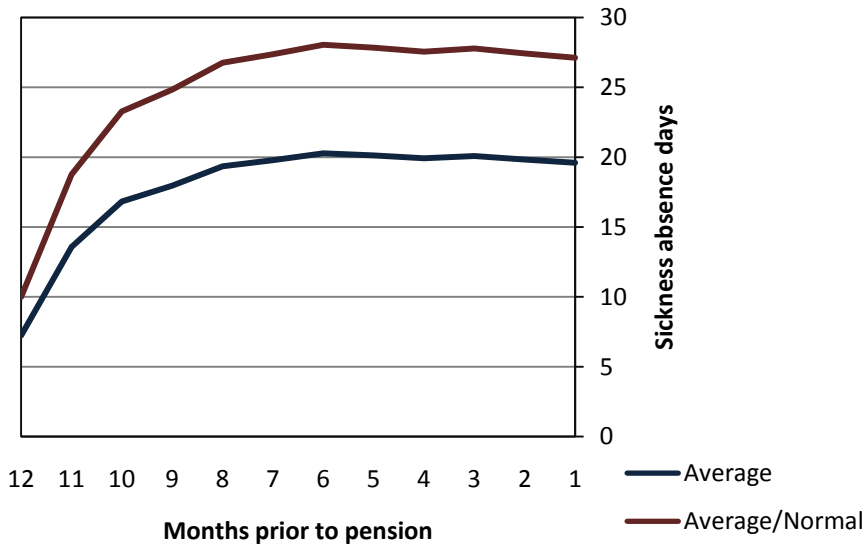

Figure 27. Average number of sickness absence days per month in progressive pattern

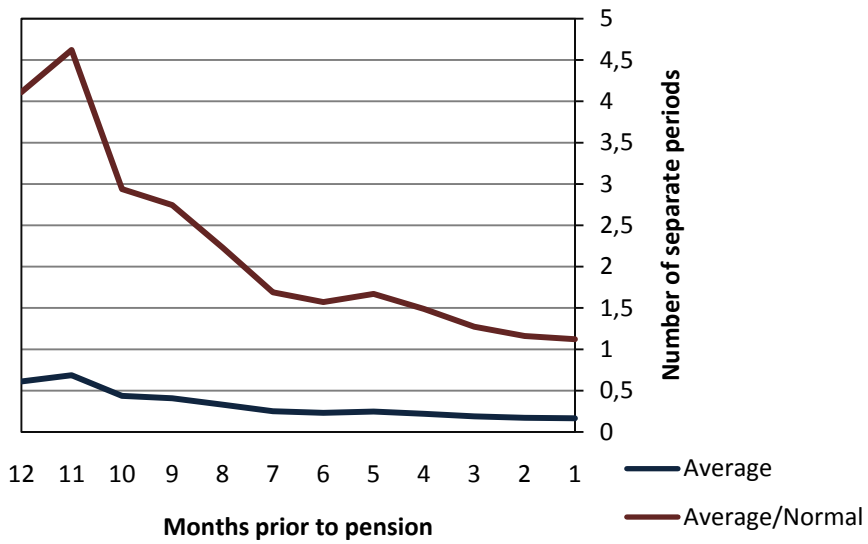

Figure 28. Average number of sickness absence periods per month in progressive pattern

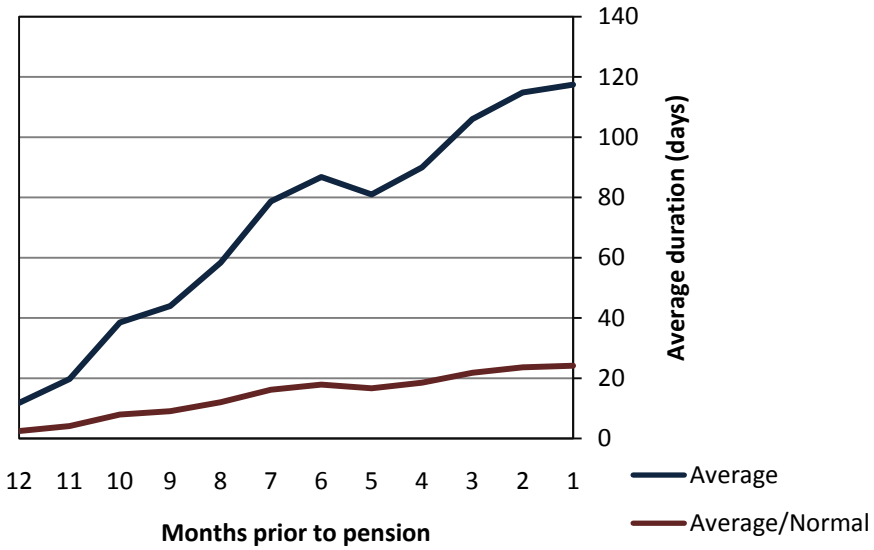

Figure 29. Average duration of sickness absences in progressive pattern

As we can see for the results, the patterns in the progressive class further reinforce the general observations made in the initial analysis. It is clear that the first sign of progressive shift to disability pension is an increased number of both sickness absences and sickness absence periods. However, as we move closer to the disability pension event, the quantity of sickness absences decreases, their length increases and the number of sickness absence days increases.

As a result we can create a progression through the following states.

1) Healthy individual (normal absence figures)
2) Frequently ill (many absence periods, many absence days, slightly above normal duration)
3) Progressive illness (few absence periods, many absence days, very long duration)
4) Disability pension

This type of a state structure of the progressive illness will be utilized later in the state space model development process.

For the sudden class of sickness absence distribution prior to disability pension, similar graphs are presented in **Figures 30**, **31** and **32**.

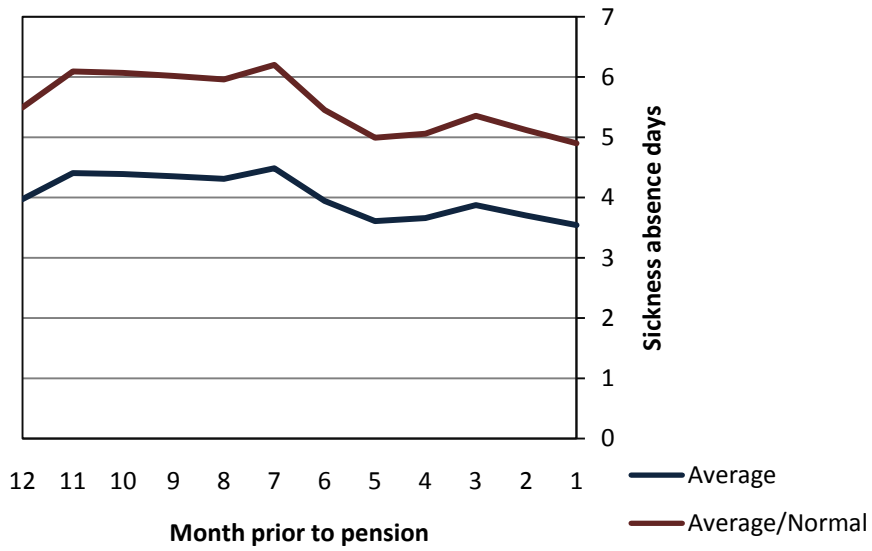

Figure 30. Average number of sickness absence days per month in sudden pattern

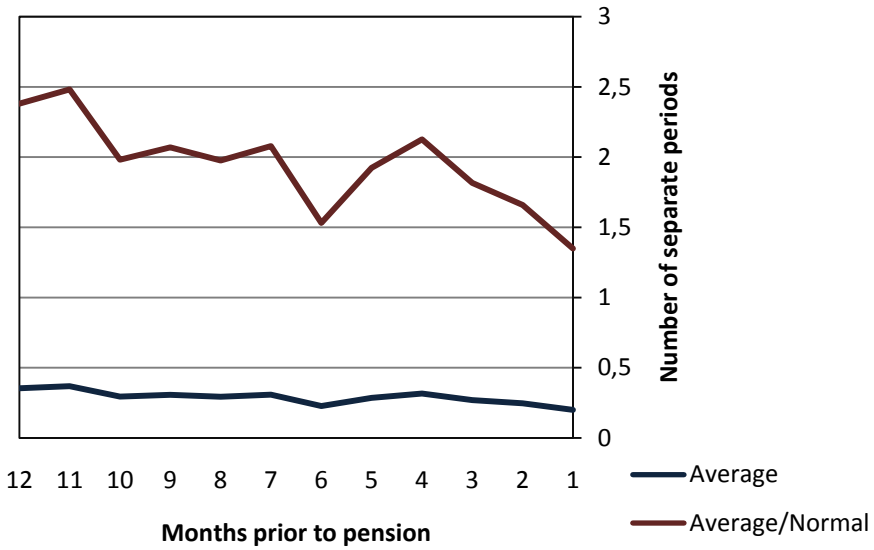

Figure 31. Average number of sickness absence periods per month in sudden pattern

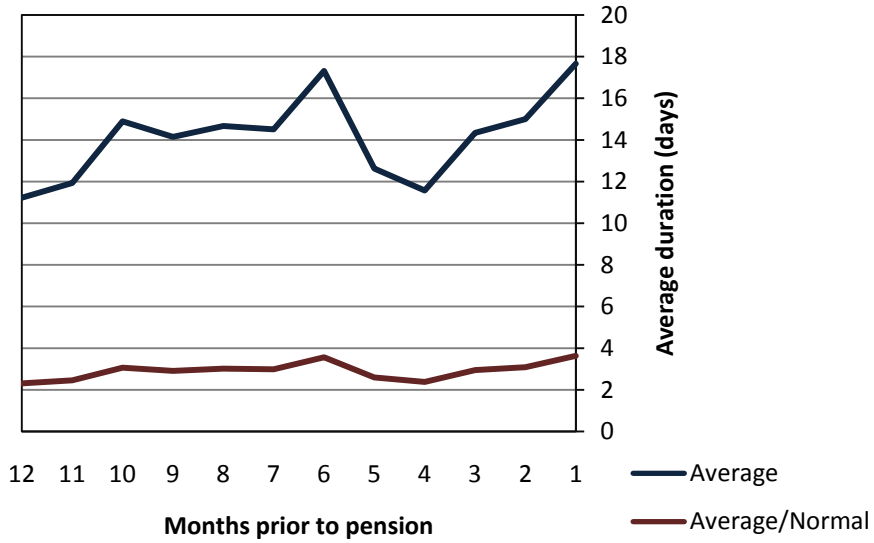

Figure 32. Average duration of sickness absences in sudden pattern

From these graphs we can see a pattern very different from that of progressive pattern. Firstly, there is no actual growth in sickness absences in this situation. The number of sickness days for the individual is on average 5-6 times the normal throughout the whole 12 months prior to disability pension. The number of sickness absence periods decreases slightly and the length increases, but these are not very significant changes. What is important is the fact that the individual seems to simply be more ill than average but does not progress through different states, but rather stays as a frequently ill individual for a long term. The states for this situation would be simpler:

1) Healthy individual (normal absence figures)
2) Frequently ill (many absence periods, many absence days, slightly above normal duration)
3) Disability pension

Having a basic understanding of the differences between the progressive and sudden sickness absence patterns prior to a disability pension we will proceed to analyze the relationships between these classes and other variables present in the data set.

3.5.2 Relationship with background variables

The core interest once again falls on their relationship to the diagnosis, gender, age and disability pension type. As a part of this analysis we perform the relevant cross-tabulations and use the χ^2-test to evaluate if there is a statistically significant effect of these background variables on the type of sickness absence pattern which will occur prior to a disability pension.

The gender effect is the simplest one to be analyzed. The hypothesis is that there should be no significant relationship between these two variables. However, the observed relationship was highly significant even at a significance level of 0,1%. The cross-tabulation is presented in **Table 6**. Male employees seem to have significantly more progressive patterns, while female employees more often have the sudden pattern. The reason here could be the fact that there is also a relationship between gender and the common type of non-accidental illness for the employee.

Table 6. Cross-tabulation of gender and sickness absence pattern

Table of Progressive by Gender				
		Gender		
		Male	Female	Total
Pattern				
Sudden	**Frequency**	249	464	713
	Col Pct	**43.15**	**55.70**	
Progressive	**Frequency**	328	369	697
	Col Pct	**56.85**	**44.30**	
Total	**Frequency**	577	833	1410

The relationship with age was more interesting. The relationship was significant at 0,1% significance level. The cross-tabulation is presented in **Table 7**. As it can be seen from the

table below, the younger individuals seem to be more likely to move to disability pension through the sudden sickness absence pattern, while the older individuals, who were born in the 1940's and are nearing the retirement age are more likely to have progressive sickness absence pattern.

Table 7. Cross-tabulation of birth decade and sickness absence pattern

Table of Progressive by Birth Decade							
		Birth Decade					
		1940	1950	1960	1970	1980	Total
Pattern							
Sudden	**Frequency**	257	326	97	29	4	713
	Col Pct	**44.23**	**54.97**	**53.89**	**59.18**	**57.14**	
Progressive	**Frequency**	324	267	83	20	3	697
	Col Pct	**55.77**	**45.03**	**46.11**	**40.82**	**42.86**	
Total	**Frequency**	581	593	180	49	7	1410

The relationship with disability pension diagnosis was also significant at 1% significance level. The corresponding cross-tabulation is presented in **Table 8**.

Table 8. Cross-tabulation of diagnosis and sickness absence pattern

Table of progressive by Diagnosis						
		Diagnosis				
		1MT	2VK	3TU	4MU	Total
Pattern						
Sudden	**Frequency**	239	45	232	197	713
	Col Pct	**49.08**	**38.79**	**57.28**	**49.00**	
Progressive	**Frequency**	248	71	173	205	697
	Col Pct	**50.92**	**61.21**	**42.72**	**51.00**	
Total	**Frequency**	487	116	405	402	1410

The results are not surprising. The noticeable deviations are the circulatory illnesses, where they are very likely to have a progressive sickness absence pattern and the disability related to limbs, where they are more likely to have a sudden sickness absence pattern.

Finally, we look at the relationship with pension type, which is significant at 0,1% significance level. The cross-tabulation is presented in **Table 9**.

Table 9. Cross-tabulation of pension type and sickness absence pattern

Table of progressive by Pension Type							
		Pension Type					Total
		8	9	S	Y	Z	
Pattern							
Sudden	**Frequency**	75	159	164	2	313	713
	Col Pct	**93.75**	**37.59**	**28.42**	**66.67**	**95.72**	
Progressive	**Frequency**	5	264	413	1	14	697
	Col Pct	**6.25**	**62.41**	**71.58**	**33.33**	**4.28**	
Total	**Frequency**	80	423	577	3	327	1410

We can see that the pension type is fairly strongly related to the type of sickness absence pattern of the individual. The full disability pension with rehabilitation is almost fully granted to individuals with the sudden sickness absence pattern, while the partial rehabilitation support and full disability pension with no rehabilitation is given more often to those with progressive sickness absence pattern. The reason for this could be the fact that during the progressive growth of sickness absences rehabilitation was already attempted with no success. Additionally, the partial permanent disability pension seems to be given almost solely to the individuals with sudden sickness absence pattern. This is most likely linked to the fact that partial disability pensions are granted to individuals who have lost part of their working capacity, but there are no predictions on further progression of the illness. This could be related to certain accidents or certain illness, which does not develop further. This also means that it is less likely that the illness has been progressing prior to the disability pension event. Additionally, if the person receives partial disability pension, then it also implies that he is capable of working more than an individual who has received full disability pension. This should also be visible prior to the disability pension event.

As a result of the analysis in this section we have obtained a general understanding of two distinguishable sickness absence patterns which may occur prior to a disability pension. These two patterns have also distinct state structure and relationship to the outcome event. A general summary of the two patterns is presented in **Table 10**.

Table 10. Summary of sickness absence pattern characteristics

Pattern	Sickness Absences	Pension Types	Diagnosis
Progressive	Increasing number of absence days Increasing length of absence periods	Partial rehabilitation Permanent full disability pension	Circulatory
Sudden	Stable but high number of absence days Stable but high number of absence periods Slightly longer length of absence periods	Full rehabilitation Permanent partial disability pension	Limbs

This categorization concludes and acts as the key result of the event study section. It will be later applied in the development of the state space model for predicting disability pensions in the population, where the fact that each of the patterns leads to specific pension types with higher probabilities will be utilized for forecasting.

Section summary

Main internal characteristics of sickness absence and disability pension data are established.

Event analysis of disability pension reveals two sickness absence patterns: progressive and sudden. Progressive sickness pattern involves an increasing number of absence days with increasing period length, while sudden sickness pattern is a more stable high level of absences, absence periods and their length.

4 Model development

The key goal of this thesis is to develop and evaluate a new individual level statistical model to estimate the risks of disability pensions in organizations. This model will be developed utilizing both the results from the literature study and also the observations made in the exploratory analysis of data. Additionally, it is important to benchmark the model performance and this will be done against a simple logistic regression model developed in the first part of this section. This will be followed by the development and evaluation of the state space model. Finally, the model results and the evaluation of their evaluation will be discussed in the next section.

4.1 Data modification

The model development process, like several of the previously presented analyses also requires specific adjustments to the data set in order to simplify the development and evaluation of the model. In this case the adjustments are related to the form of the data set, the way observations are defined and once again the data availability.

Firstly, as we have previously stated, having a 3 year sickness absence data set with evenly distributed disability pension events makes a 12 month history for forecasting a reasonable setup for the model. This time, however, only observations with at least a full 12 month history could be used. This limited the number of disability pension events to 976. The individuals with no disability pensions were taken as a single observation each. The latest sickness absence data available was used in the data set. In this way, the observations with no disability pensions accounted for another 96272 data points.

The sickness absence data was adjusted in a similar way as in the previous sections, where the 12 months data prior to the disability pension event (or latest 12 months data) was used. All of the other background variables were kept.

Another consideration was the development of the outcome variable. Firstly, the simplest variable which was developed was a binary variable indicating an event of a disability pension or lack thereof. However, due to the observations made on the sickness patterns it

was also decided to add separate binary variables to represent the 4 main pension types (full and partial disability pension, full and partial rehabilitation allowance).

Finally, the data was separated randomly into 2 sets. The first set, containing 58348 values (60%) was used as a data set for model specification, while the remaining values were used to test the model forecasting power.

This first data set would allow specifying a very short-term model, which would forecast the following month's disability pension risk on the basis of a single year of data. This is not a very difficult task, because of the strong relationship between the sickness absence time series and the disability pension risk very close to the actual disability pension event. Due to this reason, another data setup was developed, which would allow testing the model on a slightly longer time scale. For this model 24 months of data were used for each observation. This further reduced the number of disability pension observations with sufficient history to 521. The number of the remaining observations which did not include disability pension events did not change.

Using this second data set with 24 months of data prior to a disability pension event it would be possible to test a variety of forecasting setups. For example, a proposed method is to use 12 months of data to forecast disability pension events within the next 12 months. In this way the model will resemble 1 month model, but the transfer functions will be evaluated for a time unit of 12 months. This would closely resemble actual financial forecasting for a year and is a reasonable test for the model. This 12 month forecasting horizon will bring significantly more value than a 1 month equivalent. Naturally, this will only hold true if a sufficiently high predictive power can be reached on a 12 month forecasting horizon.

4.2 Simple logistic regression model

Firstly, to evaluate a reasonable benchmark model and to set a minimum goal which should be surpassed by a state space model, a simple logistic regression model was developed to predict disability pension events in the sample. Separate regression models will be developed to forecast the general disability pension events and then each type of the disability pension events. Due to the very strong relationship between the disability pension events and the sickness absence data, the hypothesized predictive power will be high in the case of general disability pension event prediction, but due to the lack of ability to identify disability pension

patterns in the logistic regression model, the power to predict specific classes of disability pensions will not be very high.

The first step in fitting the logistic model was the selection of relevant variables. Since the goal was to create a strong benchmark we have started with almost the full set of variables. The only variables which were dropped out were the unpaid sickness absences, since they were already in previous analysis identified to be too rare and possibly too volatile.

In the first fit, probability of a disability pension event in the following month was estimated. The coefficients of the first fit are presented in **Table 11**. The numbers after sickness absence variables indicate the number of months prior to the time at which a disability pension event or lack thereof is observed. A reference of all variable names is available in the appendix.

Table 11. Coefficients of basic logistic regression fit

	Coefficient	Standard Error	Z	P-value	95% Confidence Interval Lower	Upper
Intercept	-0,518	0,353	-1,469	0,142	-1,210	0,174
Gender	0,102	0,107	0,947	0,344	-0,109	0,312
BirthYear	-0,914	0,063	-14,534	0,000	-1,037	-0,791
Short12	0,050	0,011	4,564	0,000	0,029	0,072
Short11	0,030	0,009	3,321	0,001	0,012	0,047
Short10	0,062	0,008	7,562	0,000	0,046	0,078
Short9	0,018	0,008	2,175	0,030	0,002	0,034
Short8	0,031	0,008	3,803	0,000	0,015	0,046
Short7	0,041	0,008	5,024	0,000	0,025	0,057
Short6	0,024	0,009	2,723	0,007	0,007	0,042
Short5	0,020	0,009	2,210	0,027	0,002	0,038
Short4	0,019	0,007	2,737	0,006	0,005	0,032
Short3	0,004	0,007	0,565	0,572	-0,009	0,017
Short2	0,047	0,007	6,470	0,000	0,032	0,061
Short1	-0,004	0,003	-1,052	0,293	-0,010	0,003
Pshort12	-0,211	0,110	-1,926	0,054	-0,426	0,004
Pshort11	0,282	0,091	3,107	0,002	0,104	0,460
Pshort10	-0,215	0,101	-2,143	0,032	-0,412	-0,018
Pshort9	0,142	0,087	1,643	0,100	-0,027	0,312
Pshort8	0,143	0,092	1,551	0,121	-0,038	0,324
Pshort7	0,137	0,101	1,355	0,175	-0,061	0,334
Pshort6	0,237	0,116	2,052	0,040	0,011	0,463
Pshort5	0,334	0,104	3,213	0,001	0,130	0,537
Pshort4	0,127	0,096	1,332	0,183	-0,060	0,315
Pshort3	0,036	0,097	0,374	0,708	-0,154	0,227
Pshort2	-0,098	0,100	-0,972	0,331	-0,294	0,099
Pshort1	-0,235	0,084	-2,799	0,005	-0,399	-0,070

The model's Akaike Information Criterion (AIC) value was 3704,83. It can be seen that almost all of the short sickness absence day counts were highly significant in the regression. Additionally, the role of the birth year (in a decimal form where, for example, 1948 is translated to a value of 4,8) was very important. The number of periods of sickness absences was, on the other hand, less important. Nevertheless, the relationship to the sickness periods was similar to what was expected. Around 12 months prior to the disability pension there can be a large number of sickness absence periods, but closer to the actual disability pension event, their number decreases.

In order to be able to compare various model setups, a benchmark has to be developed. Since the core interest within the scope of this paper is accurately forecasting disability pensions, the predictive power is the key characteristic of the model. If we assume that the null hypothesis is the lack of a disability pension event, then we will analyze type I errors – false indications of disability pension and type II errors – failure to predict a disability pension. The balance between these two error types can be modified by varying the threshold value at which the regression output is classified as a disability pension. The exact desirable balancing point depends on the costs incurred in the case of each error type. This would be a comparison of rehabilitation and disability pension costs and possible externalities of these events. For example, setting this at a logical level of 0,5, the model produces the following results in **Table 12**.

Table 12. Basic model results on data set

		Actual		
		Disability	No Disability	
Predicted	Disability	231	81	312
	No Disability	366	57670	58036
		597	57751	58348

As we can see, the regression model successfully predicts 231 disability pensions out of 597 in total. At the same time, 81 individuals were marked as risky in terms of disability pension, but the event of disability pension did not take place. This is a fairly good result, but clearly the type I errors in this case are less problematic than type II, since its more important to be able to predict a disability pension rather than to overestimate the disability pension risk in some cases. Reducing the threshold value to 0,2 (obtained on the basis of several trials to achieve the most balanced outcome) yields the following result presented in **Table 13**.

Table 13. Basic model results with threshold value of 0,2 on data set

		Actual		
		Disability	No Disability	
Predicted	Disability	303	223	526
	No Disability	294	57528	57822
		597	57751	58348

Now, the model predicts around 50,7% of the disability pensions and still manages to have 57,6% of its predictions correct. This is a fairly good threshold level. When the model is tested on a validation sample, the following results are obtained. **Table 14** shows the results with 0,5 threshold value and **Table 15** with 0,2.

Table 14. Basic model results with threshold value of 0,5 on validation set

		Actual		
		Disability	No Disability	
Predicted	Disability	139	61	200
	No Disability	240	38460	38700
		379	38521	38900

Table 15. Basic model results with threshold value of 0,2 on validation set

		Actual		
		Disability	No Disability	
Predicted	Disability	192	140	332
	No Disability	187	38381	38568
		379	38521	38900

The result fits the expectations for a good model. When applied to the validation sample, the model properties do not significantly change and the model is still able to predict 50,7% of the disability pensions with 42,2% of type I errors, which is satisfactory.

Additionally it was decided to test a simpler model for 1 month forecasting. The decision was to keep only 5 months of data related to the sickness absence days and keep the individuals decade of birth. The remaining variables were dropped from the regression. The resulting output is presented in **Table 16**.

Table 16. Coefficients of reduced logistic regression fit

	Coefficient	Standard Error	Z	P-value	95% Confidence Interval Lower	Upper
Intercept	0,028	0,319	0,089	0,929	-0,597	0,654
BirthYear	-0,930	0,058	-16,082	0,000	-1,044	-0,817
Short5	0,079	0,006	12,968	0,000	0,067	0,091
Short4	0,054	0,006	8,771	0,000	0,042	0,066
Short3	0,031	0,007	4,741	0,000	0,018	0,044
Short2	0,069	0,006	11,782	0,000	0,058	0,081
Short1	-0,004	0,002	-1,924	0,054	-0,009	0,000

The results of this model on the validation data set with threshold values of 0,5 and 0,2 are presented in **Tables 17** and **18**.

Table 17. Reduced model results with threshold value of 0,5 on validation set

		Actual		
		Disability	No Disability	
Predicted	Disability	95	59	154
	No Disability	284	38462	38746
		379	38521	38900

Table 18. Reduced model results with threshold value of 0,2 on validation set

		Actual		
		Disability	No Disability	
Predicted	Disability	173	173	346
	No Disability	206	38348	38554
		379	38521	38900

It can be seen that the simplified model performs worse than the original model, but that the difference is not dramatic. The model with 0,2 threshold value now predicts 45,6% of disability pensions with 50% of type I errors. This indicates the fact that the predictive power of the logistic regression model is based mostly on the strong link between the sickness absences in the several last months prior to a disability pension.

As a result, the logistic regression model sets a first benchmark on the achievable predictive power:

Less than 50% type II errors with less than 50% type I errors when predicting general disability pensions in the following month with 12 months of data.

As mentioned previously, another important model benchmark was the ability to predict specific types of disability pensions. The original disability pension variable was replaced with a variable representing a certain pension type and a separate regression was built for each pension type. As a result, the model now forecasted the probability of occurance of a specific type of disability pension within a period of 1 month in the future.

The regression was performed with the full setup for the 4 main pension types (full and partial disability pension, full and partial rehabilitation allowance). The results for the partial rehabilitation allowance (type 9) are presented in **Table 19**. The results for the other pension types had similar form and for convenience only the performance figures will be presented.

Table 19. Coefficients of logistic regression fit for type 9 pensions

	Coefficient	Standard Error	Z	P-value	95% Confidence Interval Lower	Upper
Intercept	-8,109	0,613	-13,230	0,000	-9,311	-6,908
Gender	0,201	0,177	1,133	0,257	-0,147	0,549
BirthYear	0,155	0,089	1,746	0,081	-0,019	0,329
Short12	0,021	0,015	1,384	0,166	-0,009	0,050
Short11	0,006	0,012	0,479	0,632	-0,018	0,029
Short10	0,064	0,011	5,997	0,000	0,043	0,086
Short9	0,002	0,012	0,193	0,847	-0,020	0,025
Short8	0,036	0,011	3,197	0,001	0,014	0,058
Short7	0,059	0,011	5,392	0,000	0,037	0,080
Short6	-0,002	0,011	-0,190	0,849	-0,023	0,019
Short5	0,023	0,009	2,500	0,012	0,005	0,042
Short4	0,037	0,008	4,664	0,000	0,021	0,053
Short3	0,008	0,009	0,972	0,331	-0,008	0,025
Short2	0,042	0,010	4,402	0,000	0,023	0,060
Short1	0,005	0,005	1,041	0,298	-0,004	0,013
Pshort12	0,109	0,161	0,676	0,499	-0,206	0,424
Pshort11	0,316	0,134	2,359	0,018	0,054	0,579
Pshort10	-0,038	0,153	-0,245	0,806	-0,337	0,262
Pshort9	0,305	0,122	2,510	0,012	0,067	0,544
Pshort8	0,109	0,157	0,691	0,490	-0,199	0,417
Pshort7	-0,308	0,225	-1,369	0,171	-0,748	0,133
Pshort6	0,410	0,193	2,124	0,034	0,032	0,789
Pshort5	-0,684	0,263	-2,605	0,009	-1,199	-0,169
Pshort4	-0,040	0,196	-0,202	0,840	-0,425	0,345
Pshort3	-0,369	0,244	-1,514	0,130	-0,847	0,109

We can see that there is a lower number of significant predictors now, which is natural. The direct relationship between the pension type and each item in the time series is not trivial. The performance of the model for the partial rehabilitation allowance with the threshold value of 0,2 is described by **Table 20** (on validation sample).

Table 20. Logistic model results with threshold value of 0,5 for type 9 pensions

		Actual		
		Disability	No Disability	
Predicted	Disability	41	98	139
	No Disability	76	38685	38761
		117	38783	38900

The model is able to predict around 35% of the occurrences of partial rehabilitation allowance with 70,5% type I error. This is not a very poor result, but is already fairly difficult to use in actual decision making and risk analysis.

The performance figures for other disability pension types are listed in **Tables 21**, **22** and **23**.

Table 21. Logistic model results for full disability pension with rehabilitation support

		Actual		
		Disability	No Disability	
Predicted	Disability	0	7	7
	No Disability	30	38863	38893
		30	38870	38900

Table 22. Logistic model results for permanent full disability pension

		Actual		
		Disability	No Disability	
Predicted	Disability	81	78	159
	No Disability	63	38678	38741
		144	38756	38900

Table 23. Logistic model results for partial permanent disability pension

		Actual		
		Disability	No Disability	
Predicted	Disability	2	14	16
	No Disability	85	38799	38884
		87	38813	38900

As we can see, the logistic regression model provides a fairly good level of predictions on general disability pensions, but fails to provide reliable conclusions with regards to pensions with rehabilitation support. The reason is again within the sickness absence structures. The progressive sickness absence pattern is linked to the two types of disability pensions which are predicted well, while the sudden sickness pattern is related to the other two. This means that the logistic regression model is simply only capable of identifying the final high level of sickness absences related to the progressive sickness absence pattern identified previously.

Nevertheless, we can still set another benchmark on the achievable predictive power:

Less than 65% type II errors with less than 70% type I errors when predicting full disability pensions or partial rehabilitation allowance in the following month with 12 months of data.

Less than 90% type II errors when predicting partial disability pensions and full rehabilitation allowance in the following month with 12 months of data.

Finally, we have one more model specification remaining, namely the forecasting with a 12 month and not 1 month horizon. This means that now, we use data from months 13 to 24 prior to the disability pension event or lack thereof. This is a significantly more difficult forecasting task and for this reason especially with a logistic regression model the expected results are not high. Nevertheless, it is important to test this setup since relatively good output for the state space model in this forecasting task would be desirable.

The regression output for the new model is presented in **Table 24**.

Table 24. Coefficients of logistic regression for 12 month forecasting horizon

	Coefficient	Standard Error	Z	P-value	95% Confidence Interval Lower	Upper
Intercept	-1,893	0,233	-8,133	0,000	-2,349	-1,437
Gender	0,425	0,124	3,419	0,001	0,181	0,668
BirthYear	-0,691	0,042	-16,475	0,000	-0,773	-0,609
Short13	-0,001	0,006	-0,216	0,829	-0,013	0,011
Short14	0,027	0,011	2,368	0,018	0,005	0,049
Short15	0,026	0,010	2,623	0,009	0,007	0,045
Short16	0,028	0,011	2,656	0,008	0,007	0,049
Short17	0,021	0,010	2,208	0,027	0,002	0,040
Short18	-0,041	0,010	-3,961	0,000	-0,061	-0,021
Short19	0,043	0,014	3,107	0,002	0,016	0,071
Short20	0,030	0,012	2,600	0,009	0,007	0,053
Short21	-0,030	0,018	-1,685	0,092	-0,066	0,005
Short22	-0,004	0,018	-0,213	0,831	-0,040	0,032
Short23	0,019	0,019	0,997	0,319	-0,019	0,057
Short24	0,089	0,016	5,732	0,000	0,058	0,119
Pshort13	-0,145	0,068	-2,132	0,033	-0,279	-0,012
Pshort14	-0,008	0,084	-0,100	0,921	-0,173	0,156
Pshort15	0,147	0,087	1,692	0,091	-0,023	0,316
Pshort16	0,351	0,101	3,463	0,001	0,153	0,550
Pshort17	0,303	0,096	3,147	0,002	0,114	0,491
Pshort18	0,472	0,119	3,978	0,000	0,240	0,705
Pshort19	-0,083	0,131	-0,631	0,528	-0,340	0,174
Pshort20	-0,056	0,110	-0,505	0,613	-0,271	0,160
Pshort21	0,403	0,109	3,699	0,000	0,190	0,617
Pshort22	0,275	0,101	2,723	0,007	0,077	0,472

We can see that almost all of the sickness absence time series data comes into the model with a positive sign and is significant. As a result, the model now suffers in terms of forecasting power, because on the 12 month forecasting, unlike in the 1 month time horizon, the direct link between sickness absence numbers and the disability pension likelihood is significantly weaker. The resulting performance figures can be seen in **Table 25**.

Table 25. Logistic model results for 12 month forecasting horizon

		Actual		
		Disability	No Disability	
Predicted	Disability	5	41	46
	No Disability	203	37776	37979
		208	37817	38025

On the 12 month forecasting horizon, the model is unable to act as any type of a benchmark, because the model only predicts about 2,5% of the correct disability pension events and makes 8 times more type I errors.

To summarize the performance of the benchmark logistic regression model, we can say that is only able to incorporate some modeling of the progressive sickness absence pattern, where in the final half of a year prior to a disability pension, the number of sickness absences is very high. As a result, the logistic regression model provides a reasonable short-term forecasting capabilities, but is neither unable to forecast on a longer time horizon or forecast disability pension types related to the sudden sickness absence pattern. These are the key issues, which the use of state space model will address.

4.3 State space model

The state space model was selected as the modeling method in this thesis due to several reasons. Firstly, it gives the opportunity to incorporate theoretical considerations when specifying the model states and in this way creates a clear framework which is based on the states to which the individual belongs. Secondly, it allows not only evaluating the final events, such as disability pensions, but also the distribution of individuals over different model states such as sickness levels. This greatly increases the value of the model output in forecasting and risk management within organizations. Due to the combination of these reasons, already in Savin (2009) the state space model was selected as the key candidate for individual level modeling.

There are several important steps through which observations from the previous analysis which will be included into the state space model development. Firstly, the roles of various variables in the model will be defined in such a way that they follow the observations made in the literature analysis. After that, a fairly concrete hypothesis for the model will be developed. This model will act as a first estimation for the final state space model. It will be tested and revised in such a way that desired performance levels in terms of predictive power and model capabilities are met. This will allow us to specify the final model, which will be evaluated in further sections of this thesis.

4.3.1 Model variables

The roles of background variables about the individuals and of the sickness absence data will be separate. The core use of the sickness absence data will be the formulation of several states for the individual, which will describe employee's health. The background variables will be purely used to describe the transfer function between different states in the model.

During the model development it was also decided that the transfer function may include some or all of the variables from the sickness absence time series. The reason for this decision is a result of the estimation technique used. Due to the fact that simultaneous estimation of state specifications and transfer functions was not possible, the state specifications were based on observations rather than statistically determined conditions. As a result, the inclusion of sickness absence data into the transfer function significantly improved the model forecasting power. The cost of this decision is the loss of the Markov property of the model. As a result, all of the individuals will have a similar set of states in the model, but individuals with different background variables or different sickness absence history will have different transfer probabilities between the different states. For example, an older individual may have a higher probability associated with the transfer to a state of disability pension than a younger individual.

The model results which will be most important in the analysis are the transitions to disability pension states. These will first be looked at generally, but also the state space model will be used to forecast specific types of disability pensions. This will be done without significant adjustments to the model, because the different disability pension types will be already included into the model as different states. As a result, the events associated with moving from one state to another will be directly incorporated into the model. There will be no need to re-evaluate the model coefficients for each of the pension types like it was performed in the logistic regression model.

4.3.2 Hypothesized model

A hypothesized set of states has been developed on the basis of the literature review, the exploratory data analysis, where sickness absence patterns have been identified and several trials. The health states of the employee seemed to be most strongly based on the sickness absences and also on the changes in sickness absence numbers. Three most important states are hypothesized to have the following characteristics. The state is initially analyzed on the

basis of the sickness absence data for the previous 12 months. Naturally, it is important to note that these state specifications are not necessary statistically optimal from the perspective of the model forecasting performance. However, they are very good for interpretation due to their clear relationship to employee health state.

Healthy

Individuals who do not meet sickness absence requirements for the other states belong to the state of healthy individuals.

Frequently ill

- Approximately 4 to 15 times the normal level of sickness absence days
- Approximately 1 to 4 times the normal sickness absence periods
- Approximately 1.5 to 5 times the normal sickness absence duration
- These individuals should not fulfill requirements of the *progressively ill* state (progressively ill state has priority over this state).

These individuals are having increased sickness absences and are likely to move to disability pension with a sudden sickness absence pattern (see previous section).

Progressively ill

- More than 5 times the normal level of sickness absence days
- More than 50% increase in sickness absence days over 6 months

These individuals have progressive problems with sickness absences and are quite likely to become severely ill or move to disability pension through the progressive sickness absence pattern (see previous section).

Severely ill

- More than 15 times the normal sickness absence days
- Less than 50% increase in sickness absence days over the past 6 months

These individuals have most likely arrived at this stable state of severe illness from the *progressively ill* state. These individuals are very likely to move to disability pension.

Full disability pension

This is the state of disability pension where the individual has lost over 60% of the working capacity and is unlikely to recover. Individuals are likely to arrive at this state from *progressively ill* or *severely ill* states.

Partial disability pension

This is the state of disability pension where the individual has lost over 40% of the working capacity and is unlikely to recover. Individuals are likely to arrive at this state from *healthy* or *frequently ill* states.

Full rehabilitation allowance

This is the state of disability pension where the individual has lost over 60% of the working capacity but rehabilitation will be attempted to recover the work ability. Individuals are likely to arrive at this state from *frequently ill* state.

Partial rehabilitation allowance

This is the state of disability pension where the individual has lost over 40% of the working capacity but rehabilitation will be attempted to recover the work ability. Individuals are likely to arrive at this state from *progressively ill* or *severely ill* states.

The resulting state diagram is presented in **Figure 33**. Bold lines indicate dominating transfer probabilities.

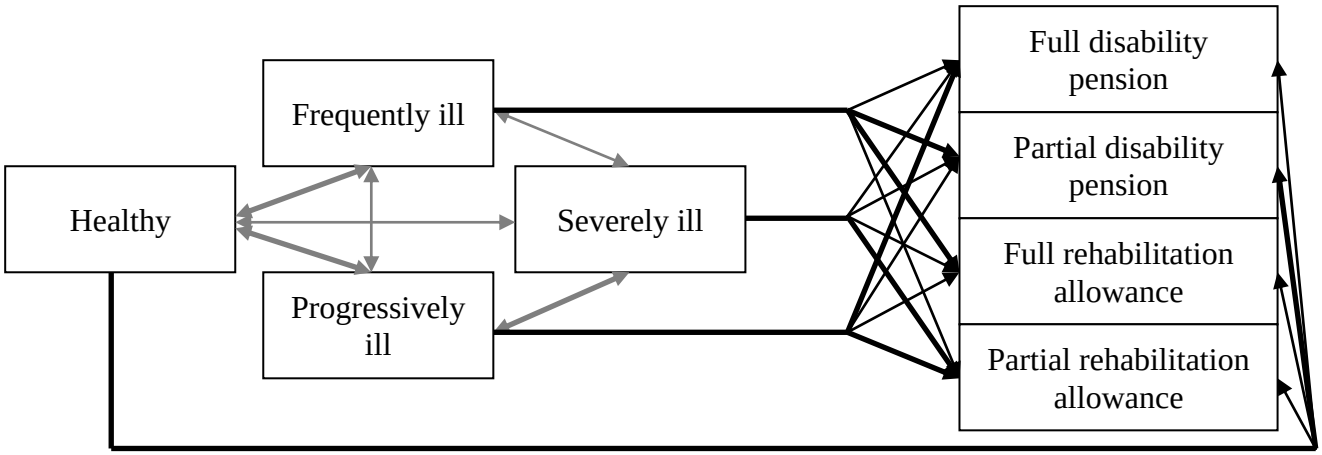

Figure 33. State space model diagram

As a result, all of the disability pension types act as terminal states in the process. The reason for this is the fact that the possible events of return from temporary disability pension are not accounted for in the data set used.

Some of the possible transition paths are also indicated with grey arrows. These are the paths which do not directly lead to a disability pension event within 1 transition. Due to the fact that all of the forecasts made within this thesis are based on a model with a single transition, the properties and estimates related to these transitions will not be covered. However, if the model will be further extended for a longer forecasting horizon, where several transitions between states are possible, these transitions will become important, because they may gradually change the health structure of the employee population.

4.3.3 Model estimation

The model estimation takes place in 2 steps. Firstly, on the basis of the historical data on sickness absences and state definitions, each of the individuals in the data set must be classified to a certain state in the model. This will require the creation of a new state variable. After this, using a suitable regression method, the coefficients for transfer functions between states will be developed in such way to maximize the predictive power of the model.

This, naturally, is a challenge from the perspective of optimization. After all, both the state definitions and the transfer functions affect the predictive power of the model, but they both can not be optimized at the same time. This means that in our model we will have to rely on previously discussed theoretical and practical observations in the development of the states and state criteria. On the other hand, the transfer functions are statistically determined using logistic regression on data showing transitions between each pair of states. This is a fairly good trade-off between the more optimal results and a more interpretable model. The states specified manually do reflect the data structure (as discussed previously), but also maintain simplicity and good interpretability, while the transfer functions will be statistically optimized and also will not be important in the output.

In model estimation, in order to avoid overlaps between variables, redundant restrictions and lack of any state assigned to an observation, the following more simple logical model was used.

if *(average sickness absence days ≥ 15* ***and***

1.5 × average sickness days in months 7-12 ≥ average sickness days in months 1-6)

then *state = "Severely Ill"*

else if *(average sickness absence days ≥ 5* ***and***

1.5 × average sickness days in months 7-12 ≤ average sickness days in months 1-6)

then *state = "Progressively Ill"*

else if *(average sickness absence days ≥ 4)*

then *state = "Frequently Ill"*

else *state = "Healthy"*

It can be noticed that the illness states are closely related to the sickness absence patterns which were discovered in this thesis. The frequently ill state was designed to correspond to individuals with sudden sickness absence pattern. The progressively ill state, on the other hand, corresponds to the first part of the progressive pattern, while the severely ill state corresponds to the final part of the progressive pattern, which occurs just before the disability pension event.

As a result, a state was assigned to each of the individuals in the data set. The first estimation was performed with 12 months of data and 1 month forecasting horizon, so the state represents the health state of the individual 1 month prior to the analysis of a disability pension event or lack thereof.

The distribution of states in the data set was the following:

54670 healthy individuals (93,7%)

1916 frequently ill individuals (3,3%)

1328 severely ill individuals (2,3%)

434 progressively ill individuals (0,7%)

As we can see, the distribution is quite good from a subjective perspective, since more severe illness and progressive illness is rarer than frequent illness and finally there is a majority of healthy individuals.

Due to a fairly good general performance of logistic regression in the benchmark model it was selected as the model for transfer probabilities. Since the time horizon of the forecast is only 1 month, it was assumed that an individual is only able to transfer to another state once. This means that the initial state space model ignored the possibility of individuals passing through several states and arriving at the state of disability pension. As a result, the model estimation consisted of 4 separate logistic regressions, each run on a specific state of individuals.

The regression results for the 4 different categories are presented in **Tables 26**, **27**, **28** and **29**.

Table 26. Coefficients of 1 month transfer function for progressively ill individuals

					95% Confidence Interval	
	Coefficient	Standard Error	Z	P-value	Lower	Upper
Intercept	1,397	1,160	1,204	0,229	-0,877	3,670
BirthYear	-0,723	0,155	-4,678	0,000	-1,026	-0,420
Short12	0,011	0,009	1,250	0,211	-0,006	0,029
Short11	0,046	0,015	3,103	0,002	0,017	0,075
Short10	0,045	0,013	3,522	0,000	0,020	0,070
Short9	0,017	0,012	1,430	0,153	-0,006	0,040
Short8	0,022	0,013	1,706	0,088	-0,003	0,047
Short7	0,023	0,012	1,893	0,058	-0,001	0,048
Short6	0,028	0,016	1,681	0,093	-0,005	0,060
Short5	0,053	0,014	3,658	0,000	0,024	0,081
Short4	0,030	0,012	2,394	0,017	0,005	0,054
Short3	0,020	0,012	1,771	0,077	-0,002	0,043
Short2	0,034	0,012	2,854	0,004	0,011	0,058
Short1	0,012	0,010	1,231	0,219	-0,007	0,032
Days	-0,179	0,114	-1,572	0,116	-0,402	0,044
Duration	-0,051	0,010	-4,943	0,000	-0,071	-0,031

Table 27. Coefficients of 1 month transfer function for severely ill individuals

	Coefficient	Standard Error	Z	P-value	95% Confidence Interval Lower	Upper
Intercept	-5,421	0,382	-14,201	0,000	-6,169	-4,673
Short11	0,031	0,031	1,025	0,305	-0,029	0,091
Short10	-0,050	0,028	-1,764	0,078	-0,105	0,006
Short9	-0,022	0,022	-0,996	0,319	-0,065	0,021
Short8	-0,018	0,021	-0,850	0,395	-0,059	0,023
Short7	-0,013	0,020	-0,628	0,530	-0,052	0,027
Short6	-0,027	0,020	-1,390	0,165	-0,066	0,011
Short5	-0,029	0,020	-1,462	0,144	-0,067	0,010
Short4	-0,029	0,021	-1,363	0,173	-0,070	0,013
Short3	-0,039	0,020	-1,989	0,047	-0,078	-0,001
Short2	0,003	0,020	0,159	0,874	-0,036	0,042
Short1	-0,055	0,019	-2,935	0,003	-0,092	-0,018
Days	0,524	0,218	2,406	0,016	0,097	0,950
Duration	0,018	0,018	0,988	0,323	-0,018	0,054

Table 28. Coefficients of 1 month transfer function for frequently ill individuals

	Coefficient	Standard Error	Z	P-value	95% Confidence Interval Lower	Upper
Intercept	-4,605	0,369	-12,488	0,000	-5,328	-3,882
Short11	-0,048	0,019	-2,481	0,013	-0,086	-0,010
Short10	-0,026	0,015	-1,787	0,074	-0,056	0,003
Short9	-0,036	0,015	-2,461	0,014	-0,065	-0,007
Short8	-0,035	0,015	-2,351	0,019	-0,065	-0,006
Short7	-0,042	0,015	-2,757	0,006	-0,071	-0,012
Short6	-0,044	0,019	-2,303	0,021	-0,081	-0,007
Short5	-0,072	0,029	-2,499	0,012	-0,128	-0,016
Short4	-0,032	0,020	-1,627	0,104	-0,071	0,007
Short3	-0,042	0,019	-2,173	0,030	-0,080	-0,004
Short2	-0,048	0,018	-2,685	0,007	-0,084	-0,013
Short1	-0,089	0,019	-4,683	0,000	-0,127	-0,052
Days	0,706	0,141	5,026	0,000	0,431	0,982
Duration	-0,025	0,034	-0,728	0,466	-0,092	0,042

Table 29. Coefficients of 1 month transfer function for healthy individuals

	Coefficient	Standard Error	Z	P-value	95% Confidence Interval Lower	Upper
Intercept	-6,080	0,104	-58,302	0,000	-6,284	-5,876
Short11	-0,049	0,031	-1,579	0,114	-0,109	0,012
Short10	-0,052	0,028	-1,836	0,066	-0,108	0,004
Short9	-0,058	0,031	-1,857	0,063	-0,119	0,003
Short8	-0,093	0,036	-2,571	0,010	-0,164	-0,022
Short7	-0,004	0,023	-0,181	0,857	-0,048	0,040
Short6	-0,052	0,032	-1,634	0,102	-0,113	0,010
Short5	-0,035	0,027	-1,285	0,199	-0,087	0,018
Short4	-0,126	0,044	-2,883	0,004	-0,211	-0,040
Short3	-0,075	0,033	-2,260	0,024	-0,139	-0,010
Short2	-0,058	0,031	-1,897	0,058	-0,118	0,002
Short1	-0,071	0,024	-2,959	0,003	-0,118	-0,024
Days	1,181	0,220	5,380	0,000	0,751	1,612
Duration	-0,137	0,125	-1,097	0,273	-0,381	0,108

From the regression tables we can see that the coefficients in from of the variables are actually dramatically different for the different states of the system. For example, the signs before the number of sickness days and their durations actually vary, showing that for some groups the increased duration of sickness absences increases risk of disability pensions, while for others the relationship is reversed. This observation shows that the separation of the individuals in the defined classes on the basis of literature analysis and empirical observations was successful (but not necessary statistically optimal).

We proceed to analyze the predictive power of the model. **Table 30** presents the general model performance on the validation sample. Due to a fairly low number of type I errors, the threshold value was dropped to 0,17.

Table 30. State space model results

		Actual		
		Disability	No Disability	
Predicted	Disability	202	200	402
	No Disability	177	38321	38498
		379	38521	38900

The model is able to predict 10 more disability pensions than the benchmark logistic regression model reaching a level of 53%. The number of type I errors is also around 50%, which is, on the other hand, higher than in the logistic regression model. Nevertheless, the model still meets the benchmark parameters and slightly exceeds the most important benchmark of 192 successfully predicted disability pensions by the logistic regression model. Additionally, we have to note that the improvement in forecasting power also comes with a reduction in number of variables present in the regressions involved in the system.

Using the data we can also evaluate the risk levels associated with each of the states in the system. The probabilities for transfer to disability pension were the following:

Severely ill – 55,8%

Progressively ill – 6,2%

Frequently ill – 4,7%

Healthy – 3,2%

Once again the level of illness is clearly linked to the risks of disability pension. This means that unlike for the logistic regression model, we can not only analyze the outcomes, but also evaluate the structure of the population to analyze anomalies in the distribution on a sub-aggregate level, for example in specific institutions.

As we have previously mentioned, the strong link between certain sickness absence patterns and disability pension types is a promising observation, because it would allow forecasting the actual disability pension type. This is an area where the benchmark model performs fairly poorly and yet it is quite important from both forecasting and rehabilitation perspective to know how severe the expected disability pension will be. For this reason, we have performed similar model transfer function estimation for each of the pension types. Due to the fact that this created as many as 16 regression outputs (1 for each combination of a state and pension type – represented by arrows on the state space diagram), the actual regression outputs are not provided and the results are directly presented. The tables with performance figures for each of the pension types are presented below.

Table 31. State space model results for partial disability pension with rehabilitation support

		Actual		
		Disability	No Disability	
Predicted	Disability	55	113	168
	No Disability	62	38670	38732
		117	38783	38900

Table 32. State space model results for full disability pension with rehabilitation support

		Actual		
		Disability	No Disability	
Predicted	Disability	0	1	1
	No Disability	30	38869	38899
		30	38862	38900

Table 33. State space model results for permanent full disability pension

		Actual		
		Disability	No Disability	
Predicted	Disability	84	148	232
	No Disability	60	38608	38668
		144	38756	38900

Table 34. State space model results for partial permanent disability pension

		Actual		
		Disability	No Disability	
Predicted	Disability	0	8	8
	No Disability	87	38805	38892
		87	38813	38900

The results show a very similar pattern to the one which was obtained in the benchmark logistic regression model. The partial disability pensions with rehabilitation support and permanent full disability pensions are predicted fairly well and actually with lower number of type II errors in comparison to the logistic regression model. The permanent partial disability pension and full disability pension with rehabilitation are, on the other hand, not predicted at all similarly to the logistic regression model. In the determination of the pension type, the

model is able to meet the benchmarks, but is unfortunately unable to surpass them. Nevertheless, the benefits of the theoretical interpretability of results and reduced variable number still hold for the state space model.

Finally, the last model test is an increased forecasting horizon. In this situation, the logistic regression model was not able to forecast the general disability pension events accurately and expectations for the model are not very high on an individual forecasting level. Nevertheless, even forecasting a few disability pensions on such time horizon would be a good achievement.

The model setup is the same as in the logistic regression model and the model performance is described by the following tables. The first one uses a threshold value of 0,15 and creates a reasonable balance between type I and type II errors. On the other hand, at the cost of type I errors the number of predicted pensions can be significantly increased as it is shown on the second table.

Table 35. State space model results for 12 month forecasting horizon

		Actual		
		Disability	No Disability	
Predicted	Disability	7	60	67
	No Disability	201	37757	37958
		208	37817	38025

Table 36. State space model results for 12 month forecasting horizon with reduced threshold value

		Actual		
		Disability	No Disability	
Predicted	Disability	28	277	305
	No Disability	180	37540	37720
		208	37817	38025

This result is not extremely impressive, but it is nevertheless fairly interesting. The model allows us to specify a group of 305 high risk individuals out of which 28 actually go to disability pension after 12 months. This is fairly long forecasting period and these results could be very valuable in actual rehabilitation and risk measurement efforts.

Section summary

A simple logistic regression model is developed and it shows fairly good results. On the most basic 1 month forecast less than 50% of type I and type II errors are present.

The logistic regression fails to forecast some of the disability pension types and has no forecasting power on a 12 month time horizon.

A state space model is developed and it shows improved results over the logistic regression model, but in many cases the improvement is not very high. The inability to forecast some disability pension types remains, but forecasting power on a 12 month horizon is improved.

5 Discussion of Model Development Results

The model performance and obtained results were only touched upon in the previous section, where the quantitative performance was evaluated. The model was able to meet the benchmarks and provided some improvements in several of them. On the other hand, there is also a variety of qualitative considerations which has to be made and several areas which may be dwelled upon. This section will contain an overview of some of these areas and will proceed to develop a methodology for the application of the model for decision making within organizations.

5.1 Interpretation of model results

Within the scope of this thesis, two models were developed to predict individual level disability pension risks. Firstly, a simple logistic regression was performed, which showed surprisingly good results in some areas of forecasting. Additionally, a basic state space model has been developed and it was benchmarked against the logistic regression model to observe if the separation of the individuals within a population into groups would allow an improvement in the forecasting power. A structured comparison of the logistic regression and state space models is presented below.

Table 37. Comparison of logistic and state space models

	Logistic Regression	State Space Model
1 month forecasting power	good	very good
12 month forecasting power	very poor	mediocre
Distinction between pension types	mediocre	mediocre
Interpretability	poor	very good
Statistical estimation	full	partial

As we can see, in general the state space model provides slightly better results in almost all of the comparison. This is a good result, but there are still areas for improvement. For example, the forecasting power over a 12 month horizon is not very strong and the ability to identify

some of the pension types has not yet been established. These shortcomings of the model will be addressed in the following two sections.

5.1.1 Weakness in pension type prediction

It has been identified that neither the state space nor the logistic regression models are able to predict full disability pensions with partial rehabilitation and partial permanent disability pensions. The reason could be related to both the data available and the nature of these disability pensions.

The full disability pension with partial rehabilitation is simply a fairly rare disability pension type. There are only 160 observations available in the population and only 30 were present in the validation sample. After the sample was split into several states, there were an extremely low number of observations left for each state. This has most likely led to a very poor estimate of model coefficients.

The other pension type which was poorly forecasted was the partial permanent disability pension. In the case of this pension type, there is only partial loss of working capacity, but at the same time no room for rehabilitation. It is likely that this category of disability pensions is poorly forecasted due to the fact that the illnesses in question are rapid one-time events which damage the working ability but are neither progressive nor continuous. An example of this could be heart stroke, which weakens the person's working ability but is not necessary predictable on the basis of prior sickness absences, since it is not a progressive illness.

5.1.2 Weakness in 12 month forecasting

The results for the 12 month forecasting seem to be significantly weaker than for the 1 month forecast. This is natural, since there is a strong and direct link between the illness and the immediately preceding sickness absences, while on a 12 month scale there may not be such a link for many illnesses the onset of which is more sudden. As a result we are able to identify only a group of 305 individuals out of which 28 will move to disability pension.

This result may seem weak, however, from the forecasting perspective we have to understand that the model will be used on a sub-aggregate and aggregate levels and not individual levels. Looking at the results from this perspective we establish a risk group of 305 individuals who have a probability of 9,2% (28 out of 305) to enter disability pension in the following year. This is actually a group which has almost 5 times the normal disability pension risk and is

obviously a valuable group of individuals for analysis. Additionally, due to the flexibility of the threshold value for the model, the type I and type II errors can be balanced in such a way to find more risk groups with different risk levels in comparison to the population average.

Essentially, on a 12 month level, the model is still a fairly good tool on a sub-aggregate level, which is the level on which the model will be used.

5.1.3 Benefits of the state space model

Having discussed some of the challenges related to the model and having outlined their reasons, it is also important to underline some key benefits of the state space model, which was developed in this thesis. Naturally, there are clear improvements in forecasting power, however, there are additional benefits which are related to the modeling technique and model structure.

Firstly, the model was developed in such a way that the interpretability of the model is maximized. This was achieved by using theoretical constructs and specifying the model states in such a way that even an individual who is unfamiliar with the model would be capable of understanding the model output on the basis of these indicators. The quantitative indicators will be further refined into a managerial-level indicator.

Another key strength of the state space model is the ability to analyze the internal structure of the population. In this way, the model not only allows evaluating the risk of the disability pensions, but also allows viewing the health states of the employees. Analysis of such data over time would allow the decision makers to notice degradation in health of the employee population long before the actual disability pension risk starts to increase significantly.

This type of analysis is especially valuable on the sub-aggregate level where parts of the population are benchmarked against each other and against the aggregate level. This would allow identifying organizations with problems related to employee health and on the other hand the organizations which have developed best practices for health promotion in the work environment. As a result, despite the fact that targeting employees on the individual level is impossible, organizations could be targeted with pre-emptive measures and best practices could be shared between organizations.

Generally, the model benefits show that not only quantitatively but also qualitatively the state space model is a significant improvement over the logistic regression model. In order to

illustrate the applicability of the model to sub-aggregate analysis, the next section provides a case study of an anonymous organization within the data sample. Additionally, a managerial indicator is developed to visualize the population structure.

5.2 Sub-aggregate analysis

The key are of use of the developed model is the analysis of employee groups and organizations. While individual level analysis is not acceptable in actual practice, comparing certain organizations against others is an effective way to estimate the risks associated with disability pensions and also to improve the forecasting related to disability pension expenses within different organizations. For this reason the analysis of samples from the population and identification of their employee structure is likely to be the most important application of the developed model.

The data set used in this paper contains also an identification variable for specific government offices and agencies which act as employers for the employees in the population under analysis. This is a very interesting variable, because it allows establishing a linkage between employees and organizations and in this way performing analysis on a sub-aggregate level. In this section, the developed model will be used to perform this type of analysis and an indicator will be developed to illustrate the internal sample structure.

5.2.1 Population and sample parameters

In order to mask the identity of the organization under analysis, the whole population was first split into two parts. After this individuals belonging to the given organization from one part of the population were used as the sample for analysis. Essentially, a random set of individuals from the given organization was taken.

Within the whole population part, the following distribution of states was present.

Table 38. Population statistics

	Share	Disability pension risk
Severely Ill	0,74 %	35,51 %
Progressively Ill	2,12 %	7,52 %
Frequently Ill	3,46 %	4,13 %
Healthy	93,69 %	0,26 %

These values will be taken as the normal values for the population. The sample used contained 2201 individuals.

5.2.2 Model estimations

The state space model with the previously estimated coefficients was run on the given organization sample of 2201 individuals. The state distribution and the corresponding estimated disability pension risk conditional on the state was estimated. The results are presented in the following table.

Table 39. Sample statistics

	Share	Disability pension risk
Severely Ill	1,14 %	31,88 %
Progressively Ill	1,32 %	11,78 %
Frequently Ill	3,23 %	15,83 %
Healthy	94,32 %	0,97 %

We can see that the distribution slightly differs from the population distribution. Comparing the sample estimation and observed population figures we obtain the following table, where the values indicate the ratio of observed to normal.

Table 40. Sample statistics in comparison to population

	Share	Disability pension risk
Severely Ill	153,95 %	89,78 %
Progressively Ill	62,20 %	156,72 %
Frequently Ill	93,30 %	382,85 %
Healthy	100,68 %	370,70 %

Using this estimation as a first step, the sub-aggregate structure can be analyzed in depth. Already these figures allow us to observe several deviations in the sample from the population. Firstly, there seems to be significantly more severely ill employees, while the number of progressively and frequently ill employees is low. This indicates general problems with large numbers of sickness absences in the organization. On the other hand, the healthy, frequently ill and progressively ill employees in this organization tend to leave to disability pension much more often than in the population.

From a subjective perspective, these observations could mean several things. Firstly, the organization could have a poor rehabilitation policy and strict sickness absence requirements, which pushes up the disability pension risks for employees who do not take a very high number of sickness absences. On the other hand, it could also mean that the reporting for short sickness absences is weak and only longer ones are actually recorded.

5.2.3 Visualized indicator

It is clear that the data presented in the table above is valuable, however, it may also be difficult to interpret without understanding the calculation procedure. For this reason, visualization was developed to present the same sub-aggregate analysis results in a more accessible form.

In the visualization the different states are placed in growing order of severity from healthy to severely ill individuals. Also, the associated colors indicate the severity level. In the normal case each of the states is represented by a rectangle of equal width. If the share of a given state is larger than normal, it receives a wider rectangle, if it is smaller, the rectangle is narrower. Each rectangle is split into two shares, the lower one representing the disability pension risk. If the risk is above normal, the shaded lower are is large. If the risk is below normal, the shaded area is low. The solid top part of each rectangle in this way shows the recovery rate from a given level of illness. Finally, the actual employee numbers and the predicted pension numbers are given below the diagram.

In this way, the diagrammatic representation of the analysis results is concise and does not create significant loss of information. The diagram below shows this visualization for the data sample discussed earlier.

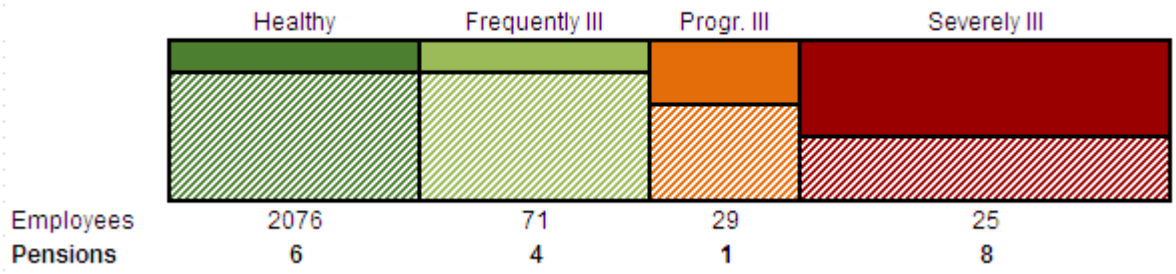

Figure 34. Visual sample statistics

From the diagram, the same conclusions can be made as the ones which were made from the data table. It is immediately clear that the disability pension rates in the 2 left categories are much too high, while the recovery rate from severe illness is surprisingly good. On the other hand, the amount of severely ill employees is very high.

Apart from the general observations, concrete pension predictions are given. In this way, the diagram both gives a general impression and a concrete result from the perspective of financial planning and employee health monitoring. This is exactly the type of information, which can be used and acted upon by the Finnish State Treasury and the government offices and agencies under analysis within the scope of this thesis.

Section summary

The shortcomings of the model are explained by the data set and by the nature of certain disability pension types.

There is a variety of qualitative benefits of the state space model, such as the improved interpretability and ability to use it for sub-aggregate analysis.

The analysis of organizations within the population is a powerful tool, its capabilities are demonstrated on an example and a visual indicator, which simplifies the interpretation of results, is presented.

6 Summary and Conclusion

This thesis acted as a continuation to the research performed by Savin (2009) with the goal of developing a model which could be used on individual level data and evaluating its predictive power. Acknowledging the role of behavioral and psychological factors in the decisions related to absenteeism and early retirement and the lack of theoretical analysis of these factors in Savin (2009) this paper started with an in-depth analysis of empirical and theoretical research in this area. Additionally, relevant and most recent literature concerning the relationship under analysis was reviewed.

Keeping the theoretical frameworks in mind, an exploratory analysis of the data set provided by Finnish State Treasury and Ministry of Finance was performed and two distinctive sickness absence patterns were identified. It is the combination of this observation and of the key takeaways from the literature analysis, which allowed the formulation of a state structure for the state space model. The state space model, which was the main output of this research managed to meet the most important benchmark criteria on the individual level. The model's forecasting power on a 1 month forecasting horizon was very strong, while on a 12 month forecasting horizon the model was able to predict a part of disability pensions, but started to suffer from type I errors. Nevertheless, this was a good result, considering the complexity and often the unpredictability of the phenomenon under study. Especially the qualitative benefits, such as the improved interpretability and the applicability of the model to analysis of organizations within the data set were the benefits which were perceived to be more valuable from the perspective of actual model application in financial and managerial decision making.

The model application is especially highlighted on the sub-aggregate level, where a methodology for model application within organizations was developed and a visual indicator was created to convey the model results in an even more accessible way in comparison to the raw numerical output. The ability to identify the anomalies within the state structure and the conditional disability pension risks is one of the most valuable outputs the model can provide.

6.1 Further research and model development

The model developed within the scope of this thesis is only the beginning of a longer project at the Finnish State Treasury, which will focus on the improvement of the ability to forecast

disability pensions in government offices and agencies. The current model supports the existence of the relationship between disability pensions and historical sickness absence data analyzed by Savin (2009) on the individual level. It also provides a concrete model implementation, which provides good forecasting performance on a short-term time horizon.

The main focus of further model development within the project of the Finnish State Treasury is the refinement of the model and the improvement of the forecasting power on a medium time horizon of at least 1 year. This type of predictions would be significantly more beneficial from the perspective of financial planning within government offices and agencies. One of the limitations of the current model was the availability of only 3 years of sickness absence history. The expansion of this period could allow for an improved model specification with longer-term forecasting power.

Additionally, the state specifications of the current model can be further developed. The current model used empirical observations and theoretical frameworks for the process of state space specification. On the other hand, additional input by actual decision makers and possibly specialists in employee health and wellbeing would allow development of more practical and realistic employee health states. Finally, specification of criteria for the classification of employees to specific states could be developed in such a way that the estimation of these criteria is performed statistically, which would optimize the model performance.

Finally, this thesis brings an additional insight into the background variables affecting the disability pension risk and the relationship between sickness absence patterns and disability pension types. The study of sickness absence patterns has allowed gaining significant insight into the model development process and this seems to be an area, which can be applied further in the analysis of employee wellbeing and the relationship between sickness absence patterns and more concrete types of illnesses. Further research in this area would be valuable not only in the field of early retirement, but could also allow to automatically triggering specific types of rehabilitation treatment to employees. This type of efficiently targeted treatment may allow decreasing the probability of early retirement and in this way increase the valuable working population, which is especially relevant under the current demographic conditions.

6.2 Conclusion

The goal of this thesis was to analyze individual level sickness absence data for a fairly large population and to develop a forecasting model based on a state-space system, which was analyzed and described in Savin (2009). Throughout the research process these goals were continuously addressed and the output was gradually tailored to meet the initial goals and expectations.

The research goals defined early on in this project were not only met, but also exceeded. The most notable additional branch of the research process was the development of a more concrete managerial indicator and model application methodology for organizations. The inclusion of this additional area was a result of an appropriate state space specification, which has allowed for simple and effective interpretation of model results, which underlines the practical value of the developed model.

From the perspective of academic research, the paper provides new empirical support for the existence of a statistical relationship between disability pensions and sickness absences on a individual level. Additionally two distinct sickness absence patterns are identified. The identification of these highly different sickness absence patterns may aid analysis of sickness absences in a variety of other academic fields and practical applications.

From the applied perspective, the developed state space model meets the basic benchmark requirements and offers not only a tool to analyze disability pension risk estimates, but also an opportunity to study the internal structure of the employee population at various organizations.

In this way, this paper has managed to integrate a strong theoretical framework, empirical data analysis and practical interpretation to create convincing support for further research in the area of the analysis of workplace health and early retirement.

Section summary

This thesis answers the initial research question and delivers a practically valuable model, which can be used for financial forecasting, rehabilitation and analysis of employee health distribution at organizations.

This thesis also advances academic research by providing support for a personal-level statistical relationship between sickness absences and disability pensions. Additionally, valuable observations about sickness absence patterns are made.

7 References

Bolin, K., Eklöf, M., Hallberg, D., Höjgård, S., Lindgren, B. (2008), Early Retirement, *Contributions to Economic Analysis,* vol.285

Bratberg, E., Gjesdal, S., Mæland, J., G. (2009), Sickness absence with psychiatric diagnoses: Individual and contextual predictors of permanent disability, *Health and Place,* vol.15

Burton, W., N., Schulz, A., B., Chen, C.Y., Edington, D., W. (2008), The association of worker productivity and mental health: a review of the literature, *International Journal of Workplace Health Management,* 1(2)

Finnish State Treasury (2009), State pension statistics 2008

Friis, K., Ekholm, O., Hundrup, Y., A. (2008), The relationship between lifestyle, working environment, socio-demographic factors and expulsion from the labour market due to disability pension among nurses, *Scandinavian Journal of Caring Studies*, 22(2)

Howarth, J. (2005), Absence management, *Strategic Vision*, 21(9)

Hosmer, D., W., Lemeshow, S. (2000), *Applied Logistic Regression, 2nd edition*

James, P., Cunningham, I., Dibben, P. (2006), Job retention and return to work of ill and injured workers, *Employee Relations*, 28(3)

Kaiser, P.O., Matsson, B., Marklund, S., Wimo, A. (2008), Health and disability pension – an intersection of disease, psychosocial stress and gender. Long term follow up of persons with impairment of the loco motor system, *Work*, 31(2)

Lee, S., Blake, H., Lloyd, S., The price is right: making workplace wellness financially sustainable, *International Journal of Workplace and Health Management*, 3(1)

Muir, J. (1994), Dealing with Sickness Absence, *Work Study*, 43(5)

Muir, J. (1983), Ill Health at Work, *Industrial Management & Data Systems*, 83(3/4)

Newton, R., Ormerod, M., Thomas, P., Disabled people's experiences in the workplace environment in England, *Equal Opportunities International*, 26(6)

Nurminen, M., M., Heathcote, C., R., Davis, B., A., Puza, B., D. (2005), Working life expectancies: the case of Finland 1980-2006, *Journal Of The Royal Statistical Society Series A*, 168(3)

Rhodes, S., Steers, R. (1981), A Systematic Approach to Diagnosing Employee Absenteeism, *Employee Relations*, 3(2)

Savin, M., (2009), Sickness Absences as an Indicator of Disability Pension Risk, BSc thesis at Helsinki School of Economics

Vanden-Eijnden, E., Stochastic Calculus lecture support material, http://www.cims.nyu.edu/~eve2/chap3.pdf, accessed on 08.04.2010

Väänänen, A., Kumpulainen, R., Kevin, M., V., Ala-Mursula, L., Kouvonen, A., Kivimäki, M., Toivanen, M., Linna, A., Vahtera, J. (2008), Work-Family Characteristics as Determinants of Sickness Absence: A Large-Scale Cohort Study of Three Occupational Grades, Journal of Occupational Health Psychology, 13(2)

Wallman, T., Wedel, H., Palmer, E., Rosengren, A., Johansson, S., Eriksson, H., Svärdsudd, K. (2009), Sick-leave track record and other potential predictors of a disability pension. A population based study of 8,218 men and women followed for 16 years, *BMC Public Health*, vol.9

8 Appendix

8.1 Key variables from the data set used in the current study

The following variables were present for all individuals:

Variable	Symbol	Description
Gender	Gender	The gender of the individual
Birth year	Birth Year	Variable which is equal to the birth year minus 1900 divided by 10. As a result 1949 will translate into 4,9.
Agency code	n/a	An identification number for the employer organization
Employment relationship	n/a	An identification number for the position at an organization
Sickness absence type	n/a	Separates paid and unpaid sickness absences
Sickness absence data	Short# Pshort# Long# PLong#	There are three variables for each of the months where the total number of sickness absences in working days, the number of sickness absence periods and their average length are recorded

The following variables are present only for individuals with disability pension

Starting date of disability pension

Disability pension size in Euros

Disability pension type

This variable distinguishes different types of disability pensions. The possible values are:

- *Full disability pension with rehabilitation support (type 8)*
- *Disability pension with partial rehabilitation support (type 9)*
- *Permanent full disability pension (type S)*
- *Partial permanent disability pension (type Z)*
- *Personal early retirement scheme (type Y)*

Diagnosis

This variable defines the primary cause for disability pension. The possible values are:

- *Mental illness (1MT)*
- *Illness of the circulatory system (2VK)*
- *Illness or disability related to moving limbs (3TU)*
- *Other diagnoses (4MU)*