Towards a data-driven circular economy: predicting material streams in the construction industry

Joel Himanen

School of Science

Thesis submitted for examination for the degree of Master of Science in Technology.

Helsinki 9.12.2022

Supervisor

Prof. Alexander Jung

Advisors

D.Sc. (Tech.) Sami Jokela

Prof. Marko Turpeinen

Aalto University
School of Science

Aalto University
School of Science

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the master's thesis

Author Joel Himanen

Title Towards a data-driven circular economy: predicting material streams in the construction industry

Degree programme Computer, Communication and Information Sciences

Major Machine Learning, Data Science and Artificial Intelligence **Code of major** SCI3044

Supervisor Prof. Alexander Jung

Advisors D.Sc. (Tech.) Sami Jokela, Prof. Marko Turpeinen

Date 9.12.2022 **Number of pages** 41 **Language** English

Abstract

The arguably inevitable transition from linear, "take-make-dispose" economies to more circular economy models is happening with an accelerating pace on a global scale. Nevertheless, this transition is still in an early phase. Societies and industries are struggling on multiple different fronts with looping material disposal streams efficiently back to the earlier stages of their value chain. Digitalization and data-driven tools and strategies have the potential to be a vital part of the solutions that will support us through this transition.

This thesis sets out to explore the relationship between digitalization and the circular economy transition. It takes the Finnish construction industry as an example environment of this transition, and suggests ways to solve the problem of aligning supply with demand in the market of non-virgin construction materials by predicting construction and demolition waste streams. This is achieved by developing two distinct, machine learning based proof of concept models (Random forest and Bayesian linear regression) that use general target building information as features to predict the amount of concrete released from demolishing the building in question. The purpose of these models is only to show the feasability of machine learning approaches to this problem.

Both models are trained and tested on a minimal, close to realistic dataset obtained from pre-demolition audits of an actual demolition site. Using the coefficient of determination as a predictive performance metric, both models significantly outperform a baseline of predicting the data mean in every case. Moreover, the Random forest approach seems to outperform Bayesian linear regression. These results suggest that machine learning models indeed are a feasible approach to predict the material streams in question. However, data on actual realized waste amounts in demolition projects is not available in the amounts required to train a production level model. Hence, aggregating detailed demolition reports and pre-demolition audits containing also temporal distributions of material streams in a nation-wide, collective register is recommended as a point of action to increase the data-readiness of the construction industry.

Keywords Circular economy, Construction and demolition waste, Machine learning, Random forest, Bayesian linear regression

Aalto University
School of Science

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Diplomityön tiivistelmä

Tekijä Joel Himanen

Työn nimi Kohti datavetoista kiertotaloutta: rakennusalan materiaalivirtojen ennustaminen

Koulutusohjelma Computer, Communication and Information Sciences

Pääaine Machine Learning, Data Science and Artificial Intelligence **Pääaineen koodi** SCI3044

Työn valvoja Prof. Alexander Jung

Työn ohjaajat TkT Sami Jokela, Prof. Marko Turpeinen

Päivämäärä 9.12.2022 **Sivumäärä** 41 **Kieli** Englanti

Tiivistelmä

Nykyisin jo väistämättömänä pidetty siirtyminen lineaaritalouksista kiertotalouden malleihin on käynnissä maailmanlaajuisesti, ja sen vauhti on kiihtyvä. Tämä siirtymä on kuitenkin vasta varhaisessa vaiheessa. Yhteiskunnat ja toimialat kamppailevat useilla eri rintamilla kytkeäkseen materiaalien hävikkivirrat takaisin niiden arvoketjujen aiempiin vaiheisiin mahdollisimman tehokkaasti. Tämän siirtymän tukemisessa ja vauhdittamisessa digitalisaatio, sekä datavetoiset työkalut ja strategiat ovat eittämättä ratkaisevassa roolissa.

Tämän diplomityön tarkoituksena on tutkia digitalisaation ja kiertotaloussiirtymän välistä suhdetta. Työssä tarkastellaan suomalaista rakennusteollisuutta esimerkkiympäristönä ja esitetään purkujätevirtojen ennustamista ratkaisuksi erityisesti ei-neitseellisten rakennusmateriaalien kysynnän ja tarjonnan kohtaanto-ongelmiin. Esitys perustellaan kehittämällä kaksi erilaista, koneoppimiseen perustuvaa konseptitodistusmallia (satunnaismetsä ja Bayesilainen lineaarinen regressio), jotka käyttävät yleisiä kohderakennustietoja selittävinä muuttujina k.o. rakennuksen purkamisesta vapautuvan betonin määrän ennustamiseen. Näiden mallien tarkoitus ei ole olla suoraan sovellettavissa, vaan osoittaa koneoppimispohjaisten mallien olevan lähtökohtaisesti käyttökelpoisia tämänkaltaisten ongelmien ratkaisemiseen.

Molemmat mallit koulutettiin ja testattiin hyvin pienikokoisella, lähestulkoon realistisella datajoukolla, joka on johdettu oikeista purkukartoituksista. Käyttämällä determinaatiokerrointa ennustustarkkuuden mittarina todettiin molempien mallien ennustuskyvyn olevan merkittävästi vertailutasoa (datajoukon keskiarvon ennustaminen kaikissa tapauksissa) parempi. Lisäksi todettiin satunnaismetsämenetelmän toimivan Bayesilaista lineaarista regressiota paremmin. Nämä tulokset viittaavat siihen, että koneoppimismallit ovat lähtökohtaisesti kelvollisia näiden materiaalivirtojen ennustamiseen. Tietoa purkuhankkeiden toteutumista ei kuitenkaan tällä hetkellä ole laajemmin saatavilla, puhumattakaan tuotantotasoisten mallien kouluttamiseen tarvittavista määristä. Näin ollen myös materiaalivirtojen ajalliset jakaumat sisältävien yksityiskohtaisten purkutoteumien ja -kartoitusten tuomista kollektiiviseen, maanlaajuiseen rekisteriin suositellaan toimenpiteeksi rakennusteollisuuden datavalmiuksien parantamiseksi.

Avainsanat Kiertotalous, Purkujäte, Koneoppiminen, Satunnaismetsä, Bayesilainen lineaarinen regressio

Preface

As the moment of writing this preface marks the beginning of the end of the most significant period of my life so far, I'd like to take the opportunity to express a few words of gratitude in my native language.

On etuoikeus saada tehdä diplomityö, jolla on konkreettinen, omien arvojeni kanssa linjassa oleva vaikutus yhteiskunnalliseen kehitykseen. Tästä haluan kiittää ohjaajiani Samia ja Markoa, joiden työ tarjosi puitteet jättää oma pieni kädenjälkeni kiertotalouden kehittymiseen. Kiitos myös Ari, Anna ja koko adesso Nordicsin väki, että tarjositte tarvittavan ajan ja resurssit tämän työn läpivientiin. Parempaa valmistumisen ajan työpaikkaa on hankala kuvitella.

Minulle tämän esipuheen kirjoittaminen ei merkitse ainoastaan diplomityön loppuun saattamista, vaan myös valmistautumista opiskelijaelämän kuplasta suureen maailmaan astumiseen. Kahdeksan vuoden mittainen opiskelija-aikani on muovannut minua tavoilla, joista en olisi osannut uneksiakaan. Aalto-yhteisö on ollut minulle kasvualusta, joka rohkaisee jokaista löytämään omat ideaalinsa, ja kasvamaan niitä kohti. Samoin vanhempani ovat aina antaneet minulle tilaa löytää omat polkuni. Kiitos, että unohditte kertoa, mihin en pysty.

Kiitos Fyysikkokilta, jossa sain ottaa ensiaskeleeni fuksina, teekkarina ja opiskelija-aktiivina. Kiitos Löröötí ja Fuksitoimikunta '18 unohtumattomasta kipparivuodesta, ja kiitos fuksit '18, että sain etuoikeuden olla osa teidän matkaanne teekkareiksi. Kiitos BY maailman parhaasta opiskelijan kodista, ilman teitä en olisi selvinnyt pandemia-ajasta näin vähin vaurioin. Kiitos AYY ja hallitus '20 elämäni intensiivisimmästä, mutta kirkkaasti opettavaisimmasta vuodesta. Kiitos Ströböt ja BK lukuisista reissuista, vertaistuesta ja ystävyydestä. Kiitos Hankkijat, että elämä ja kellottaminen ei aina ollut niin vakavaa. Lopuksi, kiitos Veikko Sompa ja hänen ajo-oppilaansa, että jaksoitte vaikeinakin hetkinä muistuttaa MINUA! siitä, kuka ajaa.

Usko parempaan huomiseen ja valmius tarttua toimeen sen saavuttamiseksi ovat minulle lähtemätön osa teekkariutta. Olen yhä vahvemmin sitä mieltä, että tuon idealismin kanavoiminen teoiksi on diplomi-insinöörin velvollisuus. *Iuvenis in aeternum.*

Helsinki, 9.12.2022

Joel Himanen

Contents

Abbreviations

API	Application programming interface
BLR	Bayesian linear regression
CDW	Construction and demolition waste
CE	Circular economy
GBC	Green Building Council
GHG	Greenhouse gas
HP	Hyperparameter
LOOCV	Leave-one-out cross-validation
ML	Machine learning
MVG	Multivariate Gaussian
PoC	Proof of Concept
RF	Random forest

Acknowledgements

In addition to the advisors and supervisor, the author is indebted to the following people and organizations for providing valuable insights into the world of construction and demolition:

- Elli Kinnunen, *A-Insinöörit*
- Aki Finér, *Motiva*
- Ella Lahtinen, *Green Building Council Finland*
- Harri Hakaste, *Finnish Ministry of the Environment*
- Anni Tyni, *City of Helsinki*

1 Introduction

Circular economy (CE) models are a key factor in how we change the foundations of our societal and economical structures in order to fight environmental crises, and make life on our planet sustainable. Taking the streams of material disposal and looping them around to feed back into the value chain is one of the most promising ways to decouple growth from resource constraints, and thus reduce virgin natural resource extraction to a sustainable level. [23] The construction industry, being one of the most significant CO_2 emitters, is an interesting example of a field, where the transition to more circular operating models is just starting to take shape. It is also a field that is on the more conservative side from the point of view of digitalization. That is to say, that e.g. data-driven strategies and operating models have not yet spread through the industry, as they have done in numerous others. This implies a significant unused potential, that could be a decisive factor in supporting the industry through the transitions to come.

This thesis has been written as part of the *Platform Ecosystem for Material Circulations* project of Motiva Services Ltd. through the author's employment. It attempts to clarify ways in which the construction industry could harness the yet untapped potential of data, focusing on the supply of non-virgin materials salvaged from construction and demolition waste (CDW), and specifically the problem of aligning that supply with the demand. A key challenge is estimating CDW streams before a demolition project commences, which is why the problem is approached with machine learning (ML) models, guided by the following research questions:

- *Is ML-based predictive modeling feasible in the context of CDW streams?*
- *What does the successful use of ML require from the construction industry?*

The thesis builds the necessary background in Chapter 2 by describing the high-level situation and forces that affect the global CE transition, working its way down to the practicalities of the Finnish construction industry, and finally formulating the machine learning problem that is the main focus of this work. Next, two distinct proof of concept models, a Random forest and Bayesian linear regression, are developed (Chapter 3), tested, and compared (Chapter 4). Furthermore, Chapter 5 takes the model testing results as a basis and formulates suggestions on how the industry should go about aggregating, storing, and utilizing data to support its circular economy transition. Additionally, the use of data-driven strategies to accelerate this transition is further motivated by discussing the trade of non-virgin materials from the point of view of the market economy and potential business cases. As a conclusion, critical top-level remarks are summarized in Chapter 6.

2 Background

This chapter establishes the reasoning behind the urgent need for a global transition towards a circular economy and describes the role digitalization plays in the process, building the case for developing data-driven solutions to accelerate it. It takes the Finnish construction industry as an example of this transition, and specifies the key challenges and opportunities of driving the change with data and digital solutions, also introducing the material circulations ecosystem project of Motiva Services Ltd. this thesis contributes to. Finally, it describes the concept of a pre-demolition audit, and defines a supervised machine learning problem that attempts to support that process in making more accurate predictions of released CDW streams.

2.1 From linear to circular economies

It is no news that the ongoing environmental crises, such as global warming and loss of biodiversity, pose a large part of the most significant threats of our lifetime. The challenge of steering humankind towards a sustainable way of life is monumental, and requires extraordinay feats, as well as cross-disciplinary coordination and co-operation from all entities ranging from individual people to businesses, industries, countries and international alliances.

In 2020, the Finnish Innovation Fund Sitra conducted a study on the megatrends for the beginning decade [4], identifying five different, but intertwined global megatrends, three of which are the following:

- *The urgent need for ecological reconstruction*
- *The redefinition of the economy*
- *Technology is embedded in everything*

Getting the interactions between technology and economy to work in favour of fighting environmental crises is an incredibly complex, yet critical challenge. In other words, it is safe to state that the two latter trends play a key role in the way we drive ecological reconstruction in the years to come.

The culprits behind our environmental crises are manifold, but the problem largely condenses down to how we use the natural resources our planet provides us with. Already in 2019, resource extraction and processing made up approximately half of global greenhouse gas (GHG) emissions, in addition to more than 90% of biodiversity loss and water stress. [15] It has become evident, that the sun has set on the "take-make-dispose" linear economy model. We can not afford to extract, use and dispose of resources with the pace we have reached due to assumptions of limitless economic growth with little to no regard to external costs.

The concept of a circular economy (CE) presents an alternative model to this, where the use of raw materials and production emissions are decreased by e.g. maximizing product lifetime and reusability, minimizing production waste, and implementing business models based on sharing, reusing and recycling products. [10] Figure 1 summarizes how this kind of circularity is achieved by looping the disposal streams of a linear economy model back to the different stages of the value chain. It is also important to notice the prioritization of those loops; ideally, the inner loops are the most significant ones, since they require less effort and resources to return to the user. For example, it is generally more desirable to return a product to the provider to be reused as such, than to deconstruct the product, recycle the materials and produce new parts and products out of them. The distinction between reuse and recycling is also crucial later, when observing material circulation in the construction industry more closely.

The potential in this shift towards CE is enormous: according to a Material Economics study from 2018 [6], the transition to a more circular economy could reduce GHG emissions from heavy industry in the EU by 56% by 2050. Contrary to a common misconception, this shift does not need to happen at the cost of economic growth as a whole. Aligning technologies and business models to tackle wasteful

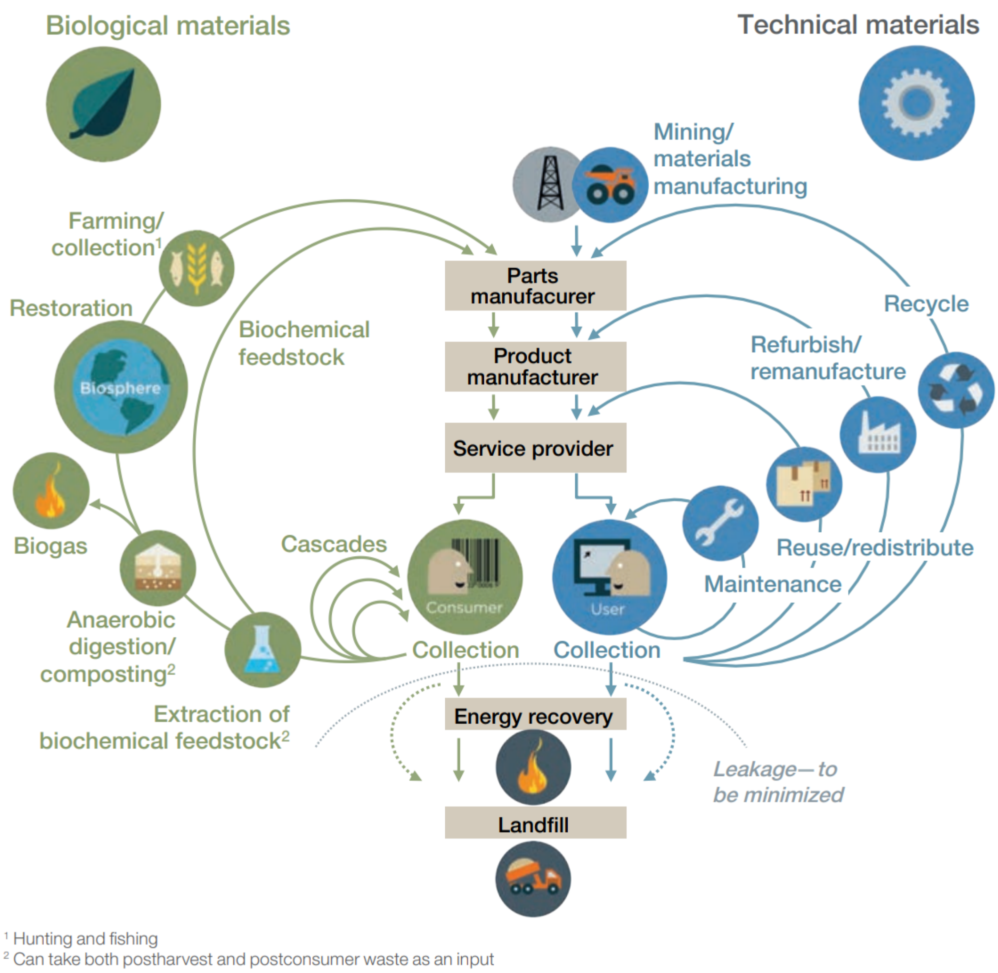


Figure 1: The various loops of a circular economy. *Source: World Economic Forum* [23]

value creation chains and shift towards a more circular model would actually create new markets, and boost employment, lifting EU's GDP by 7% and yielding annual benefits of up to €1,8 trillion by 2030. [7]

2.2 The role of digitalization and data

As Sitra states, technology is embedded in everything, and the transition to CE is no exception. The accelerating pace digitalization has developed with during the last decades has presented us with countless possibilities and opportunities.

Advances in the areas of data and digitalization, however, have come with a significant set of challenges, as well. In addition to a plethora of privacy issues related to emerging data economies, with increasing amounts of digital solutions and services increases also the demand for energy. Hence, it would be naïve to assume digitalization to be a "silver bullet" that directly solves all of our problems. This is also the case in implementing CE models, as is stated in a 2019 study by the European Policy Centre (EPC). [10] That being said, the study focuses on establishing the role of digitalization as a key enabler of CE - as long as it is properly managed. They divide approaches to use digitalization to accelerate the transition to CE into three categories:

1. Improving knowledge, connections and information sharing
2. Making business models, products and processes more circular
3. Strengthening the roles of citizens and consumers

Within the scope of this thesis, we will mainly focus on categories 1 and 2, exploring the ways proper data aggregation and utilization through ML models can support businesses in adapting more circular models. Specifically, we will look at the construction industry, as it is responsible for almost half of global raw material extraction and a quarter of global GHG emissions. [5]

2.3 Circularity in the Finnish construction industry

In order to further clarify the context of this thesis, it is necessary to gain some understanding of the different entities that play a role in the Finnish construction industry, and specifically, what their responsibilities and impacts in terms of construction material circulation is. Figure 2 illustrates how different stakeholders interact in a demolition project. This is, of course, only an example setting and a simplification, but serves to explain the general flow. In some cases, the real estate owner might assume more responsibility and drop the main contractor as a redundant link in the chain. Alternatively, sometimes larger construction companies purchase the target building and act as owner and main contractor. Often it is also the case, that the demolition contractor does the pre-demolition audit as well.

From the point of view of material flow, though, it is imperative to notice that not all materials released during demolition are the responsibility of a single entity. Reusable components and movables, that are possible to extract without heavy machinery, belong to the owner or main contractor, and have to be removed before actual demolition. This process is usually outsourced to recycling operators that

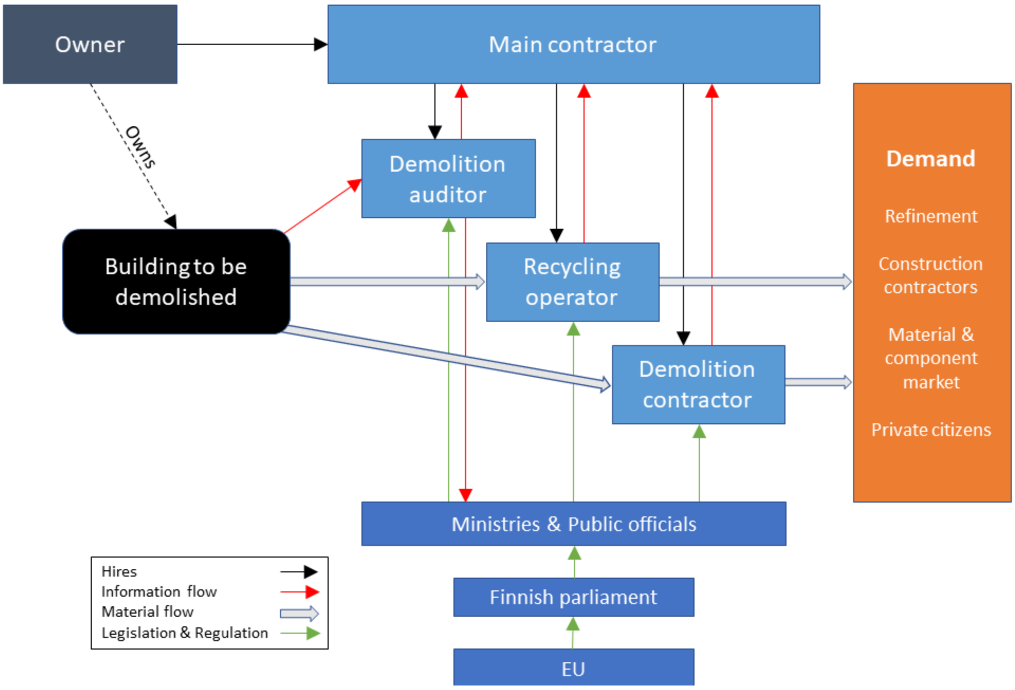


Figure 2: The different roles and their interactions in a generic demolition project.

extract the components and sell them on to new users. Reusable components, that require heavy machinery for extraction, on the other hand, are the responsibility of the demolition contractor, and have to be explicitly specified by the owner/main contractor.

When reviewing CO_2 emissions as part of the environmental impact of the construction industry, it is essential to divide them into three categories according to the phases in the lifecycle of a building:

1. **Upfront carbon** (materials production and construction phase)
2. **Operational carbon** (use phase, energy consumption)
3. **End of life carbon** (demolition phase, demolition waste handling)

These are defined by the World Green Building Council (WorldGBC) in their 2019 report "Bringing embodied carbon upfront" [1]. The same study finds that with the recent advances in the fied of renewable energy, operational carbon emissions are reduced significantly. Consequently, upfront and end of life carbon will grow quickly in proportion. In particular, the WorldGBC predicts that upfront carbon will be responsible for up to half of the whole carbon footprint of new construction between the the present and 2050. A similar trend can be seen in Finland, as well. Up until 2020, approximately 80% of the Finnish built environment's total lifecycle emissions came from operational carbon. [12] With low-carbon energy gaining foothold in the Finnish energy market, focus is shifting increasingly towards reducing upfront and end of life carbon by transitioning to a more circular operating model. The Finnish government has recently taken the lead in driving this transition across industries in Finland. In the spring of 2021, they established a strategic programme [16] to pave the way towards a new Finnish economy based on CE by 2035. The programme dedicates a significant amount of focus on the construction industry, and specifies several action points to promote CE in its context, including financial incentives, digitalization and data management.

From a CE point of view, one of the the main challenges in the Finnish construction industry is construction and demolition waste management. Legislation and various costs seem to form a setting, where there is no clear and desirable way for the responsible party to handle CDW. On a demolition site, for example, the prime contractor owns all of the materials and waste (although the responsibility of demolition waste handling lies with the demolition contractor, if the demolition is outsourced by the prime contractor), and therefore is also responsible for processing all of the material streams released during demolition accordingly, including all related costs and taxes. Current legislation makes it difficult, and in many cases expensive, to arbitrarily dispose of CDW. Obviously, this option is also undesirable from an ecological point of view, due to which it is also safe to assume that legislation is going to get even more strict in this matter. Enormous costs and the lack of proper real estate, on the other hand, make it impractical to store CDW even for short spans of time. These matters lead to the conclusion that there is also a significant cost case in addition to the ecological case, that speaks for a circular system where CDW could be salvaged immediately from site for reuse (some components can be extracted pre-demolition and used again as such) and recycling (waste that is not feasible for reuse is processed into new products).

There are, however, significant challenges to overcome. Plenty of costs are embedded in the process of preparing CDW from extraction to reuse and recycling, and said costs often outweigh the costs of disposing of this waste, which makes it less appealing for contractors to salvage it. This problem tends to have the effect of making virgin materials more attractive price-wise, which hinders the demand for reusable and recycled CDW, and makes adapting a more circular model difficult from the point of view of procurement as well. Not only that, but the whole paradigm of construction and most of its processes are built around using virgin materials; a buyer generally risks lower product quality when dealing with reused and recycled materials. In addition, companies working in the construction industry often have neither the capabilities, nor the skills required to deal with non-virgin materials and components. This problematic is clearly a key obstacle, as it effects negatively both the supply and demand sides of the non-virgin construction material market. This has also been recognized by the Finnish government, as it has recently set in motion investigation projects to e.g. identify ambiguities in construction product qualification procedures. [24] As an example result, the Finnish Ministry of the Environment released a policy brief [25] on verifying the feasability of reused materials in June 2022. It states that it is sufficient to check for feasability with construction site -specific verification, which means that a CE-marking is not required for e.g. demolition waste that is reused as such.

However, even if there was a strong financial incentive to favor non-virgin materials, a significant challenge remains aligning demand with supply in the reusable and recycled CDW market on a temporal level. Currently, contractors are able to release information of available materials only after they become available, as the demolition project proceeds. Because storage possibilities for these materials are few and costly, the time window to find a buyer is far too narrow. Additionally, this unpredictable batch-like supply leads to discontinuous availability, rather than a continuous supply of non-virgin materials. To counter this challenge, the industry will first and foremost need dedicated operators working on providing this continuous supply. On the other hand, it will also require the digital infrastructure necessary to uphold these operating models. Addressing this need, Finnish officials are currently working on a series of databases and registers that would together provide a centralized data storage system for information on demolition projects and plans, construction product characteristics and material streams, to name a few. These systems are planned to come with diverse APIs to enable private entities to develop their own services and platforms to utilize all of the data stored in these public systems.

2.4 Motiva Services Ltd. and the ecosystem for material circulations

In order to make the raw data from the existing and upcoming registers for construction materials and demolition waste available and usable to companies and consumers, it is vital that different services are developed to refine and process the data. Even though work on the state-wide registers is still in progress, the Finnish state-owned company Motiva, or more accurately, its subsidiary Motiva Services Ltd. is already working on an ecosystem and platform for construction material circulation; a project, under which this thesis work is also carried out.

Motiva is a company of experts specialized in providing businesses and municipalities with information, solutions and services on how to use energy and materials more efficiently and sustainably. With this ecosystem and platform, it is Motiva Services' aim to build a hub that collects information from various sources under a single platform, acts as a meeting place for supply and demand of non-virgin construction materials and supports all relevant stakeholders in adapting to a more circular operating model in the construction industry.

At the moment, this platform is still in the planning phase. Motiva Services has developed the concept for it, and it has been validated and further developed using two different projects as case studies. One is a demolition project that takes place in Vattuniemi, Helsinki, where 16 buildings are set to be demolished, starting in 2023. The other site is located in the Tampere area, where especially the reuse and circulation of ashes as a construction product in infrastructure building was explored. The latter project ended in Spring 2022 for the time being. These projects are vital in gathering field experience on the challenges and requirements of material circulation. Specifically, Motiva Services seeks to pilot planned features of their platform, and to gain knowledge on how exactly predictive, data-driven demolition planning supports a circular economy.

2.5 The possibilities and challenges of pre-demolition audits

A critical part of shifting towards predictive demolition planning are the procedures that guide the mapping of the demolition site. These include mapping of hazardous, but also reusable and recyclable materials that the building to be demolished consists of. Historically, only the mapping of asbestos and other harmful materials has been required by law, but nowadays it is increasingly common to do a broader material mapping. The most extensive mapping process is the pre-demolition audit (in Finnish "purkukartoitus"), which also includes, for example, reusable movables and recommendations on how to to reuse, recycle or dispose of the material streams released during demolition. [13] These audits are at the moment voluntary, but officials are constantly planning ways to encourage real estate owners and contractors to use them as a tool to increase material circulation. This is slightly challenging, since a more extensive audit is also more expensive to the client. It is, however, crucial from the point of view of CE, that pre-demolition audits are carried out as carefully as possible, because gaining knowledge of future material streams is necessary in order to align demand with supply in the non-virgin material market.

Carrying out pre-demolition audits, and especially estimating the amounts of different materials released, is hard to do accurately. All experts interviewed for this thesis agreed on this, and stated several possible reasons for the significant errors that occur even in state-of-the-art audits. Some of the material amounts are especially hard to predict, because they might often be hidden from plain sight in the foundations of the building in question. An example of this kind of material is mineral wool, which is often used as an insulator. Especially in the case of old buildings, where the quality of material documentation might be suboptimal, these kind of "hidden" materials can be nigh unto impossible to predict. Extensive, on-site analyses might yield better results, but these are mostly impractical, because the target buildings are often still in use at the time of conducting the audit. It might also happen, that some renovations or other changes during the building's lifetime have not been properly documented, which causes some material streams to come as a surprise during demolition. Finally, it is also natural that clients want to minimize the costs of a demolition project, and the pre-demolition auditors may have to compromise extensiveness and accuracy of the audit in order to stay within the budget.

Given the potential of pre-demolition audits from a CE point of view, and the nature of the challenges described above, it is evident that making the audit process more data-driven holds significant possibilities to develop construction material circulation. The obvious bottleneck at the moment, however, is the lack of data. There are no standardized, mandatory procedures to store pre-demolition audit, material mapping, or realized material stream data. This leads to the relatively small amount of feasible data to be scattered all over different stakeholders, and to have little to no quality guarantees. This challenge is at least partly being addressed by the projects described in sections 2.3 and 2.4, but the purpose of this thesis is to motivate and set a roadmap for proper data aggregation, utilization and sharing in this industry by demonstrating predictive capabilities using actual proof of concept models.

2.6 Supervised machine learning for CDW stream prediction

To counter the challenges and realize the potential of the pre-demolition auditing process described in 2.5, we now set out to solve the particular problem of predicting CDW streams to grow the window of opportunity to salvage reusable and recyclable materials from a demolition site. The approach we take on this problem is a concept that has seen rapid advances in the last decade alone: machine learning.

Being one of the most widespread modeling paradigms today, ML methods have an impact on all of our lives, whether we want them to or not. Every digital device and transaction generates data, that is utilized in increasing amounts by ML models used to e.g. personalize services and recommendations, and optimize marketing and sales.

The general concept of supervised ML is to take a set of low-level quantities (features) from a data-generating system, and make predictions of quantities of interest (labels) by mapping the features to these labels using a hypothesis map (model). An optimal hypothesis map is obtained from a typically large set of candidate maps by minimizing the discrepancy (loss) between actual label values and those predicted by the hypothesis map candidates. This procedure is referred to as "training" an ML model. [11] The word "data", in this context, refers to the combined set of features and labels.

More precisely, we will consider a supervised learning setting, where we have a dataset $\mathcal{D} = \{\boldsymbol{z}_1,\ \boldsymbol{z}_2,\ \ldots\ ,\ \boldsymbol{z}_m\}$, where a vector $\boldsymbol{z}_i$ denotes a data point consisting of features and labels, and m is the size of the dataset. The feature space $\mathcal{X}$ encloses all possible feature values, and similarly, the label space $\mathcal{Y}$ is the space of all possible label values. In the case of $\mathcal{Y}$ being discrete, we talk about classification problems. In the problem setting of this thesis however, we have $\mathcal{Y} = \mathbb{R}$, which implies a regression problem.

Let n be the number of features in the dataset. The feature vector $\boldsymbol{x} = [x_1,\ x_2,\ \ldots\ ,\ x_n]$ contains all feature values for a single data point. It is often useful to represent all feature vectors in a $m \times n$ matrix

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_m \end{bmatrix}.$$

The label of the ith data point is denoted by y_i, and by extension, all labels in the dataset are represented in the label vector $\boldsymbol{y} = [y_1,\ y_2,\ \ldots\ ,\ y_m]^T$. This notation is used for *observed* labels, while quantities *predicted* by a model are referred to using $\hat{y}$. In the scope of this thesis, we will only consider a single label at a time in order to avoid higher dimensional label matrices.

2.7 Data

The data our models are trained and tested on is based on 16 actual pre-demolition audits carried out on target buildings on the Vattuniemi demo site (2.4). Each building being represented as a data point, the set of features is obtained from initial knowledge of the building. The label set, on the other hand, includes estimates of four different materials released as CDW during demolition, calculated by domain experts in the pre-demolition audit. Some of the feature and label titles, along with an example data point, can be seen below.

Area	Volume	Storeys	Concrete	Mineral wool	Metal	Glass
9590	32400	6	9028.1	20.5	214	25.1

Here, the three chosen features are area, volume, and number of storeys, the units of the former two being m^2 and m^3, respectively. The four labels are measured in metric tons. In the test runs, the label of choice was concrete, as it is by far the most significant CDW stream. There are a number of alternative features available, such as categorical variables indicating building type or load bearing material, year of construction, and binary features indicating, whether the target building in question is equipped with an elevator or sewer system. These were, however, left out of this analysis for the sake of computational convenience, as well as the fact that there is little to no variance within the categorical features in this minute dataset.

3 The Models

This chapter introduces two distinct modeling approaches, the Random forest (RF) regressor and Bayesian linear regression (BLR), to predict demolition material streams using initial knowledge of the target building. It is imperative to note, that these models are only a proof of concept - they are meant to demonstrate that given proper data, it is possible to use various machine learning algorithms to make feasible predictions on different material streams, and thus support the pre-demolition audit process. Therefore, these models as such should not be relied on in actual cases.

Due to the exceptionally small amount of data, it was difficult to build a strong case to support any particular modeling approach, apart from keeping the model simple and the number of features low. [11] Hence, the model choices may seem somewhat random. There is, however, one principle that guided the choice of modeling methods. It is the dissimilarity of the two models, which helps to gain a more broad picture of which approaches could be feasible in this context.

3.1 Random forest regression

Introduced by Breiman (2001) [3], the Random forest algorithm is a supervised ensemble learning method, that is made up by of a group of decision trees. Decision trees and Random forests have remained a popular modeling approach because they imitate human reasoning and decision making, and are thus conceptually intuitive. They can be used for classification and regression tasks alike, but we will focus on the latter because of the nature of our problem setting.

3.1.1 Decision trees

A Decision Tree regressor is a predictor that maps features to the label space by essentially going through a series of if-else-statements. Each of these decision statements checks, whether or not a specific subset of features fills the decision criteria. This is done until the process arrives at a prediction. It is conventional to represent decision trees as graphs, that start at the top with a (root) decision node, and continue down to further decision nodes, until arriving at a leaf node. Here the decision nodes represent the if-else-clauses, and leaf nodes the predictions. A simple example of such a decision tree for three-dimensional data $\boldsymbol{z}_i = [x_1^{(i)},\ x_2^{(i)},\ y^{(i)}]$ is depicted in Figure 3. This tree shatters a training dataset into four subsets, one for each leaf node. It achieves this by taking one data point at a time, starting at the root node, and choosing the next edge to traverse based on whether the decision statement of the node is true or false in the case of the feature vector in question. If the chosen edge leads to a leaf node, the tree starts to process the next data point. If the next node is a decision node, the decision process is repeated. A new prediction is made similarly by passing the new feature vector through the decision nodes, until it lands on a leaf node. The predicted label is then given a value equal to, for example, the mean of the training labels in the subset associated with that leaf node.

An obvious point of interest remains the question of how to grow an optimal decision tree. There are several possibilities to do this, but we will consider as an

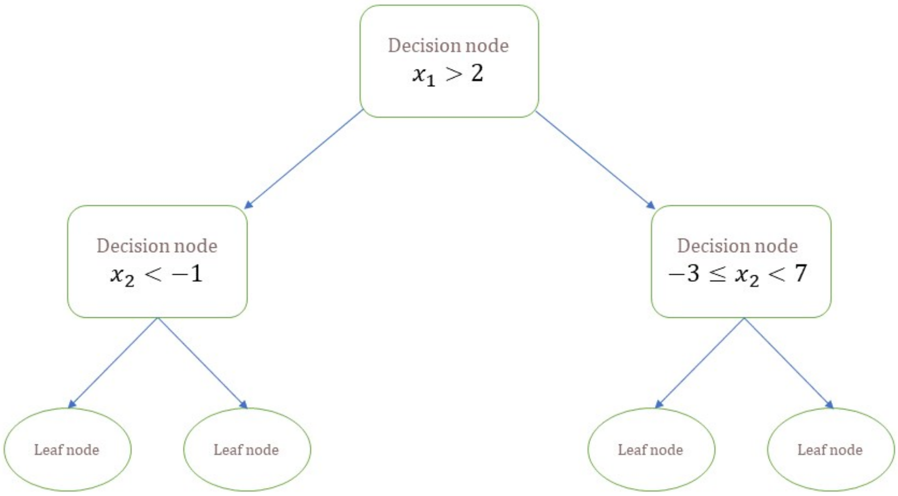


Figure 3: A graph representation of a simple decision tree.

example a variant of the ID3 (Iterative Dichotomizer 3) algorithm [18] (Algorithm 1), designed to handle regression problems.

The algorithm starts with a training set $\mathcal{D}$ and a root node. It then proceeds to recursively build the two subtrees sprouting from the root node, until maximum tree depth k is reached, or the size of the resulting subset is small enough ($< l$).

The final question to be cleared is the problem of finding the best possible splitting rule, or decision statement, at each decision node. Generally, the metric for the "goodness" of a split is called the *Gain*. Again, there are multiple possibilities when

Algorithm 1 ID3 variant for regression

max_depth $\leftarrow k$
min_samples $\leftarrow l$
procedure ID3($\mathcal{D}$, current_depth $= 0$)
  nr_samples $\leftarrow |\mathcal{D}|$
  if current_depth $<$ max_depth & nr_samples $>$ min_samples **then**
    $\mathcal{D}_1, \mathcal{D}_2 \leftarrow \arg\max_{\mathcal{D}_i, \mathcal{D}_j} \text{Gain}(\mathcal{D}, \mathcal{D}_i, \mathcal{D}_j)$
    $T_L \leftarrow$ ID3($\mathcal{D}_1$, current_depth $+1$)
    $T_R \leftarrow$ ID3($\mathcal{D}_2$, current_depth $+1$)
    return Decision node with left and right subtrees T_L and T_R
  else
    return Leaf node with set $\mathcal{D}$
  end if
end procedure

it comes to choosing a Gain function [20], but our choice will be *variance reduction*.

Let $\mathcal{D}_0$ be the dataset at the current node, and $\mathcal{D}_i, \mathcal{D}_j$ be its split, such that $\mathcal{D}_i \cap \mathcal{D}_j = \varnothing$ and $\mathcal{D}_i \cup \mathcal{D}_j = \mathcal{D}_0$. With $\mathcal{Y}$ denoting the set of labels, we can define the (weighted) variance reduction as

$$\Delta\sigma^2 = \mathrm{Var}(\mathcal{Y}_0) - \left(\frac{|\mathcal{D}_i|}{|\mathcal{D}_0|}\mathrm{Var}(\mathcal{Y}_i) + \frac{|\mathcal{D}_j|}{|\mathcal{D}_0|}\mathrm{Var}(\mathcal{Y}_j) \right). \tag{1}$$

The algorithm goes through every possible split and finally chooses

$$\mathcal{D}_1, \mathcal{D}_2 = \underset{\mathcal{D}_j, \mathcal{D}_j \subseteq \mathcal{D}_0}{\arg\max}\, \Delta\sigma^2$$

to be the respective subsets for the next nodes.

It is worth to mention, that letting the tree grow arbitrarily large results in an infinitely large Vapnik-Chervonenkis dimension for the model, and consequently in unacceptably severe overfitting. [20]

Therefore, it is critical to set the bounds k, l for tree depth and minimum split size. Optimizing these hyperparameters with more sophisticated methods than simple grid-search is out of the scope of this thesis, though.

3.1.2 Forming the ensemble

Another way to control overfitting and increase robustness of the Decision Tree model is to grow an ensemble of them and choose the average of all decision tree predictions as the final prediction. This method is called the Random forest. In order to avoid correlated decision trees in the ensemble, the Random forest algorithm trains each tree using a method called Bagging. [11]

The basic concept is to choose the training set for each tree by uniformly sampling a random subset with replacement from the original dataset. Additionally, the Random forest forces each tree to always consider only a random subset of features when choosing the best split at every decision node (line 6 of Algorithm 1). Introducing these random factors ensures that the generalization error of a Random forest converges as the number of trees in the forest increases [3], which makes it a strong tool to control overfitting.

At this point, the main weakness of the Random forest becomes evident. Computational complexity quickly becomes a problem, especially as the number of trees increases. Growing a single decision tree, especially with a large training set, is a computationally heavy task due to the need to go through all possible splits at each decision node. This effect is amplified when growing a Random forest, resulting in a theoretical computational complexity of $\mathcal{O}(k \cdot t)$ (excluding unification costs) and memory space cost of $\mathcal{O}(2^k)$, where k is the tree depth and t is the number of trees in the forest. [21]

3.2 Bayesian linear regression

The second, alternative modeling approach is Bayesian linear regression. It is arguably a slightly more advanced method, and offers the means to include domain knowledge in the model in the form of prior distributions, in addition to some metrics for uncertainty in the predictions. Given the scenario this model is applied to in this thesis, one could argue for BLR to be somewhat "overkill" of a choice. However, it goes to demonstrate that also more advanced probabilistic methods are applicable in this context.

3.2.1 A Bayesian approach to ordinary linear regression

We start with the familiar model of ordinary linear regression

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{w} + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\epsilon}$ is an error vector. Fitting this model is done by solving the optimization problem of finding the weight vector $\hat{\boldsymbol{w}}$ that minimizes the sum of squared errors (the so-called Ordinary Least Squares problem)

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w} \in \mathbb{R}^n}{\arg\min} \sum_{i=1}^{m} |y_i - \mathbf{X}\boldsymbol{w}|^2 = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y} \tag{3}$$

The Bayesian approach to this model follows the idea that we do not deal with exact values for the vectors $\boldsymbol{w}$ and $\boldsymbol{\epsilon}$, but rather place some probability distributions on them. Furthermore, we place prior distributions on the parameters of the afore mentioned distributions, which we then update with our data. The updated distribution is called *the posterior*, and it is computed with Bayes' rule

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}, \tag{4}$$

where $p(A|B)$ is the posterior (the probability of parameter A given the data B), $p(B|A)$ the likelihood (the probability of observing the data B given parameter A), and $p(A)$ is the prior on parameter A. The quantity $p(B)$ is another prior-like marginal probability, but as we will see in section 3.2.2, it can be dealt with as a normalizing constant when deriving the posterior distribution.

Obtaining the posterior can be cumbersome; at times there might not even be an analytical solution available, and one has to resort to e.g. numerical integration. There is, however, the possibility of using a conjugate prior, which makes computations more straightforward and guarantees the existence of a closed-form posterior. [19] This essentially means, that using a prior that is conjugate to the likelihood results in a posterior of the same family of distributions as the prior.

Going back to the linear model in Equation (2), we now make the assumption that our data is given rise to by a multivariate normal distribution, or Gaussian (MVG). Knowing that the conjugate prior for a MVG likelihood is also a MVG [14], we place on the weights and errors, or noise, the prior distributions

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_0, \alpha^{-1}\mathbf{I}) \tag{5}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \beta^{-1}\mathbf{I}), \tag{6}$$

with $\boldsymbol{\mu}_0, \alpha, \beta$ being hyperparameters interpreted as prior weight means, and weight and noise precision, respectively. We can now write the likelihood as

$$p(\boldsymbol{y}|\boldsymbol{w}, \beta) = \mathcal{N}(\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I}). \tag{7}$$

3.2.2 Deriving the posterior distribution

The most challenging part of BLR is finding the posterior for $\boldsymbol{w}$. Often, a closed form solution does not exist, and even given conjugacy, it still depends on the exact form of the priors and likelihood, and thus has to be derived carefully. Despite the lack of a general formula, we can still usually follow a certain routine to complete the derivation:

1. Start with Bayes' rule (4) and drop all normalizing constant terms independent of $\boldsymbol{w}$
2. Perform some algebraic manipulations
3. Due to conjugacy, recognize an unnormalized form of a distribution belonging to the same family as the prior

Following these steps, we can now derive the posterior for our case.

Using Bayes' rule (4) and our pre-defined distributions (5)-(7), we can express the (unnormalized) posterior as

$$p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y}, \alpha, \beta) \;\propto\; p(\boldsymbol{y}|\boldsymbol{w}, \beta)p(\boldsymbol{w}|\alpha) \;=\; \mathcal{N}(\boldsymbol{y}|\mathbf{X}\boldsymbol{w}, \beta^{-1}\mathbf{I})\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_0, \alpha^{-1}\mathbf{I}). \tag{8}$$

Writing out the two Gaussians, dropping constant terms and rearranging, we can state that the product in (8) is proportional to

$$\begin{aligned}
&\exp\left[-\frac{1}{2}\beta(\boldsymbol{y}-\mathbf{X}\boldsymbol{w})^T(\boldsymbol{y}-\mathbf{X}\boldsymbol{w})\right]\exp\left[-\frac{1}{2}\alpha(\boldsymbol{w}-\boldsymbol{\mu}_0)^T(\boldsymbol{w}-\boldsymbol{\mu}_0)\right] \\
&\propto \exp\left[-\frac{1}{2}\beta\boldsymbol{w}^T\mathbf{X}^T\mathbf{X}\boldsymbol{w} + \beta\boldsymbol{y}^T\mathbf{X}\boldsymbol{w} - \frac{1}{2}\alpha\boldsymbol{w}^T\boldsymbol{w} + \alpha\boldsymbol{w}^T\boldsymbol{\mu}_0\right] \\
&= \exp\left[-\frac{1}{2}\boldsymbol{w}^T\underbrace{(\beta\mathbf{X}^T\mathbf{X}+\alpha\mathbf{I})}_{\boldsymbol{\Sigma}^{-1}}\boldsymbol{w} + \underbrace{(\beta\boldsymbol{y}^T\mathbf{X}+\alpha\boldsymbol{\mu}_0^T)}_{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}}\boldsymbol{w}\right].
\end{aligned}$$

Completing the square [2], we get that the above expression is proportional to

$$\exp\left[-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{\mu})\right],$$

which we recognize as the unnormalized Gaussian $\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}(\beta\mathbf{X}^T\boldsymbol{y} + \alpha\boldsymbol{\mu}_0) \tag{9}$$
$$\boldsymbol{\Sigma} = (\beta\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}. \tag{10}$$

This concludes the derivation, as we are left with a MVG with mean (9) and covariance (10) as the posterior distribution for the regression weights $\boldsymbol{w}$.

3.3 Implementation

Both models were implemented in a Python environment, using different libraries for convenience, namely NumPy [9] for computing and pandas [22] for data processing. In the case of training and test errors, the root-mean-squared error

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2} \tag{11}$$

was used as a metric for the loss between predicted and actual values.

3.3.1 Random forest

For the Random forest model, a ready-made implementation in the Scikit-learn library [17] was used, namely `sklearn.ensemble.RandomForestRegressor`. As Algorithm 1 states, the hyperparameters characterizing the forest are maximum tree depth and minimum number of samples to split. Due to the small size of the dataset and number of features, there are not many sensible values for these hyperparameters. Still, a grid search (Algorithm 2) with the domain of integers in the interval [2, 5] for both HP's was performed in order to ensure an optimal HP configuration. Each configuration was evaluated using Leave-one-out cross-validation (LOOCV), which essentially means taking a single datapoint as a validation set, training the model on the remaining $M - 1$ datapoints, making a prediction on the validation point and computing the validation error. This is repeated for every datapoint, and the final score is the average of all validation errors. To avoid correlation between the individual decision trees, only two of the three features were randomly chosen at each split.

Algorithm 2 Grid search for hyperparameter optimization

```
Define a space of possible values for each HP
best score ← inf
for every possible HP configuration do
    Perform LOOCV
    if LOOCV score < best score then
        best score ← LOOCV score
        best HP configuration ← current HP configuration
    end if
end for
return best HP configuration
```

3.3.2 Bayesian linear regression

In contrast to the Random forest model, no existing implementations were used for BLR, but the model was built "from scratch". Regarding the hyperparameter choices described in 3.2.1, a strategy different from grid search was implemented.

Due to the small size and relatively large variance of the dataset, it is reasonable to set weakly to non-informative priors, i.e. not guide the model too much in any direction. This translates into choosing extremely small precisions α, β, or large variances for the prior distributions of the weights and errors. To reach sufficiently weak informativeness, the HP choices used were $\alpha, \beta = 10^{-4}$. Making the priors this broad also decreases the effect of $\boldsymbol{\mu}_0$, which was set to equal the zero vector $\mathbf{0}$.

4 Results

This chapter summarizes the results of training and testing both the RF and BLR models on the Vattuniemi data described in 2.7. For model performance metrics, training and test errors, as well as the coefficient of determination R^2 were used. Additionally, the difference between training and test error was observed as an indicator of overfitting.

Concerning R^2, an exact definition has to be supplied in order to avoid ambiguity. Describing how well the model in question is able to explain the data at hand, $R^2 \in [0, 1]$ is usually defined such that $R^2 = 0$ indicates a baseline equivalent to predicting the mean of the data in all cases, and $R^2 = 1$ implies a 100% accuracy, i.e. that the model is able to fully explain the data. To achieve comparability, a definition compatible with Bayesian models was used here. It was suggested by Gelman et. al. (2019) [8], and is based on prediction and residual variances:

$$R^2 = \frac{\mathrm{Var}(\hat{\boldsymbol{y}})}{\mathrm{Var}(\hat{\boldsymbol{y}}) + \mathrm{Var}(\boldsymbol{y} - \hat{\boldsymbol{y}})} \tag{12}$$

4.1 Random forest

For the Random forest model, grid search yielded a hyperparameter configuration of (3, 2), denoting maximum tree depth and minimum split size, respectively. Using the optimal HP configuration, 100 RF's with 200 trees each (the same number of trees was used in HP optimization as well) were trained and tested, using a different 80% to 20% train-test split every time. Figure 5 shows the result of six randomly selected runs, with the estimated amount of concrete plotted as a function of target building area (the other two features were ignored here due to visual convenience). Model diagnostics are depicted in Figure 6.

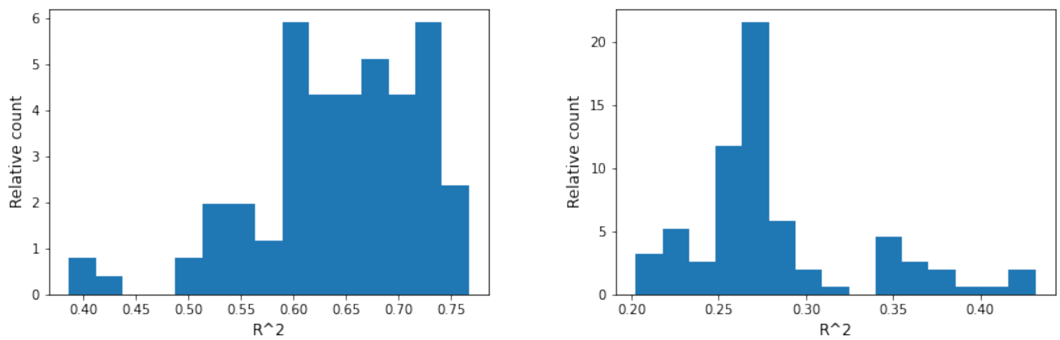
4.2 Bayesian linear regression

Due to the choices justified in 3.3.2, no HP optimization was performed here. But similarly to RF, the model was fitted on 100 different 80% to 20% train-test splits. Results and diagnostics of these test runs can be seen in Figures 7 and 8.

4.3 Comparison

A thorough discussion of the above results will follow in 5.1, but let us make a few first remarks here. It is evident from the diagnostics in Figures 6 and 8 that both models show strong variation in the results across train-test splits. In addition to that, we notice that the the test error tends to be greater than the training error for both models (6c and 8c), which indicates overfitting. [11] However, the quantity of most interest is the coefficient of determination R^2, as it can be viewed as a metric for the prediction accuracy of a fit.

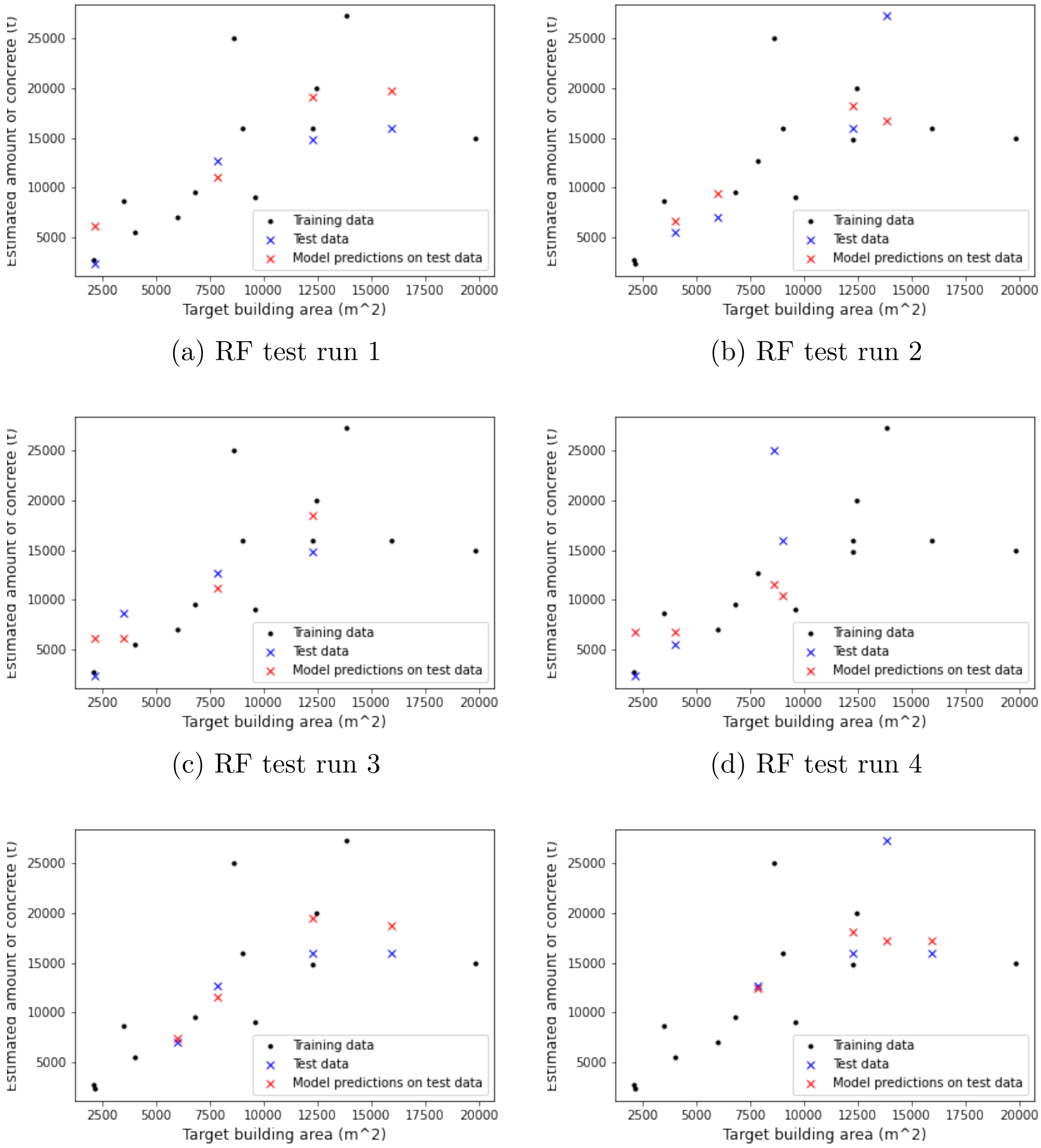
In order to be able to compare the two models on a single glance, it is useful to summarize the resulting R^2 scores of fitting both models with 100 different splits (6d and 8d) in two separate histograms. These can be observed in Figure 4. Considering that the R^2 values of the Random forests fall between 0.4 and 0.8, while the scores of the BLR fits remain between 0.2 and 0.45, it is safe to conclude, that on this dataset, the Random forest is able to explain the data better than Bayesian linear regression.



(a) R^2 histogram for RF

(b) R^2 histogram for BLR

Figure 4: R^2 comparison between RF and BLR



(a) RF test run 1

(b) RF test run 2

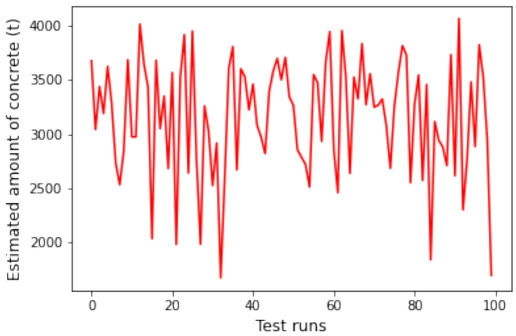
(c) RF test run 3

(d) RF test run 4

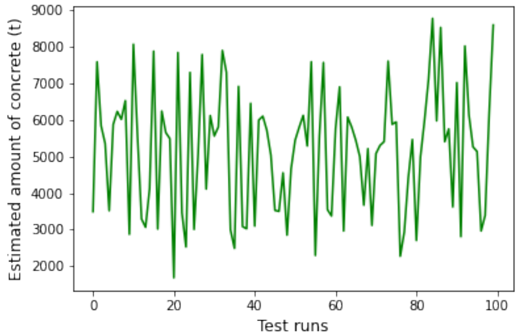
(e) RF test run 5

(f) RF test run 6

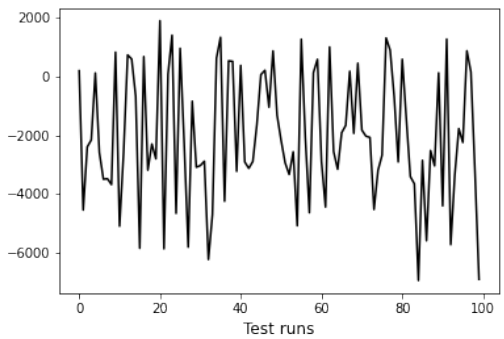
Figure 5: Six randomly selected test runs of the 100 RF models fitted on the Vattuniemi data.



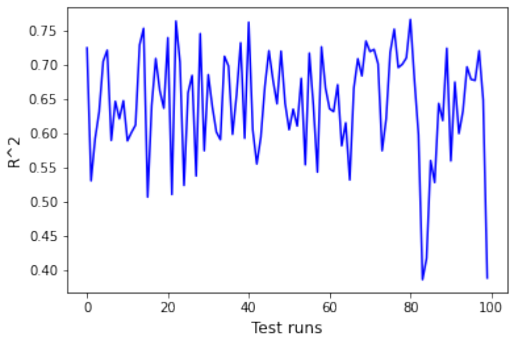
(a) Training errors



(b) Test errors

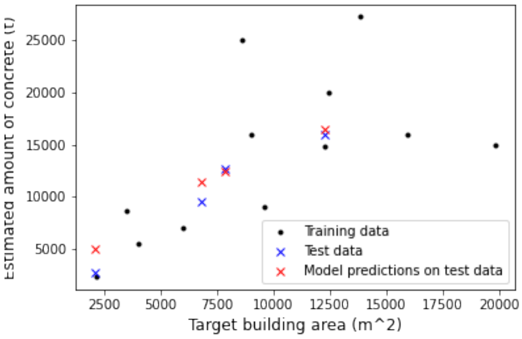


(c) Difference between training and test errors

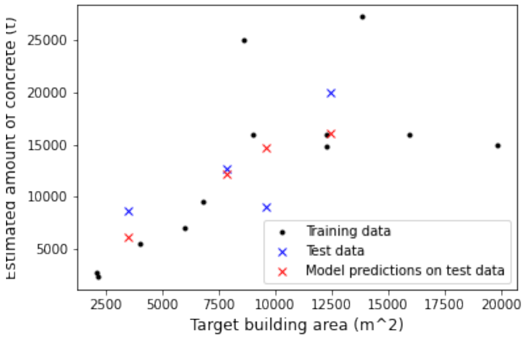


(d) Coefficient of determination R^2

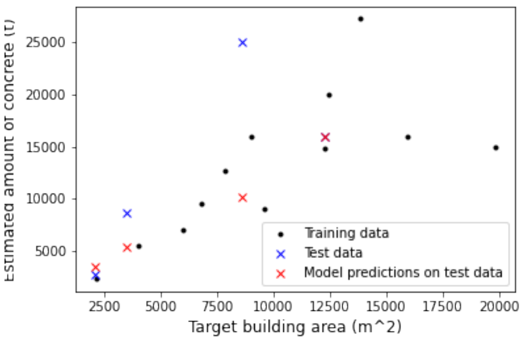
Figure 6: Model diagnostics on the 100 RF test runs



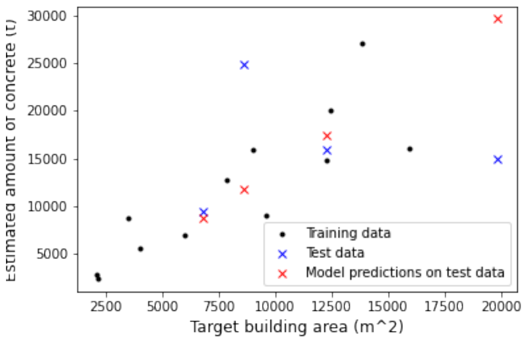
(a) BLR test run 1



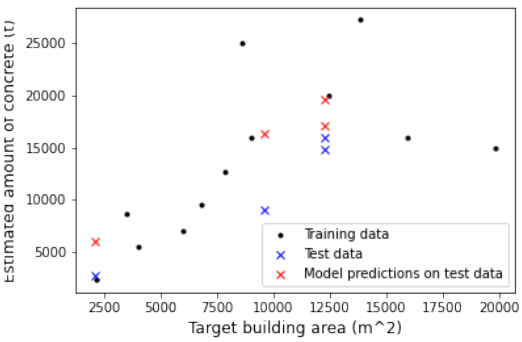
(b) BLR test run 2



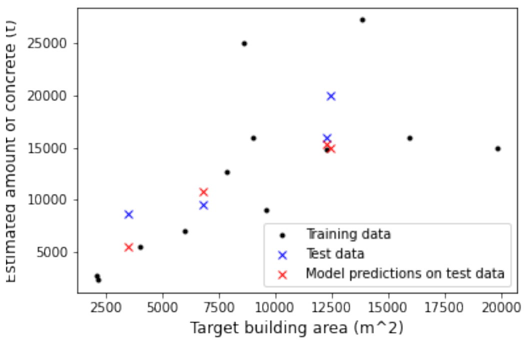
(c) BLR test run 3



(d) BLR test run 4

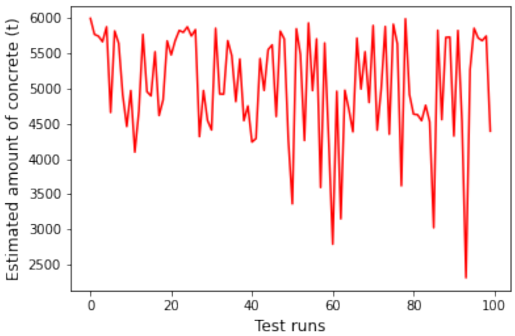


(e) BLR test run 5

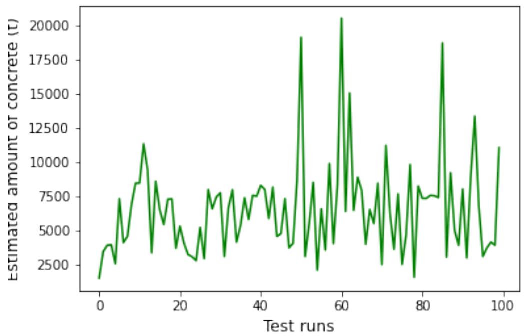


(f) BLR test run 6

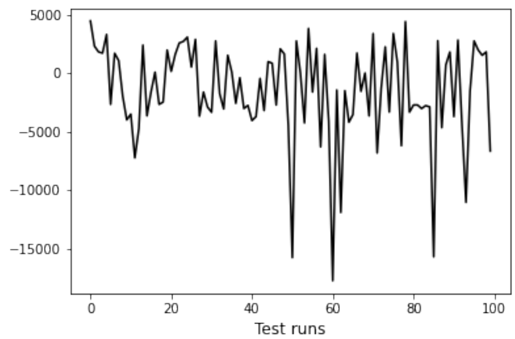
Figure 7: Six randomly selected test runs of the 100 BLR models fitted on the Vattuniemi data.



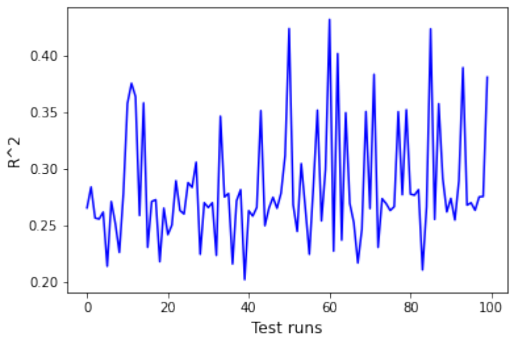
(a) Training errors



(b) Test errors



(c) Difference between training and test errors



(d) Coefficient of determination R^2

Figure 8: Model diagnostics on the 100 BLR test runs

5 Discussion

This chapter draws conclusions from the test results reported in Chapter 4, and expands on them, elaborating ways for the construction industry to utilize the full potential of its data in the transition towards a more circular economy.

5.1 Interpreting the results of PoC model testing

First, it has to be stated that this work is quite exploratory. There are no public benchmarks on using ML methods to predict CDW streams based on building traits, so expectations for these models were not high to begin with. They were intended only as PoC's, after all. On another note, the dataset we are dealing with here is exceptionally small. Only 16 datapoints are unlikely to lead to any applicable results. Even simpler models usually require significantly more data to perform reasonably well.

Looking at the test results in Figures 6 and 8, it can be concluded, that the magnitude of the test errors is too large for the models to be used as such in real scenarios. In context, the actual estimated concrete amounts range from 2000 to 30 000 tons, while the test error magnitudes rarely go under 3000, but can climb up to more than 10 000 tons. A predictive model would have to do better to be of any actual use in the pre-demolition audit process. It must also be noted that the strong variance in test errors is most likely a result of a a small testing dataset.

Another probable consequence of a small dataset is the tendency of the models to overfit. This means that the models learn to explain the training data well, but fail to generalize, i.e. to predict new datapoints accurately. This is manifested as large test errors compared to training errors, as seen in 4.3. Depending on the model in question, there are several possible ways to try and avoid overfitting. For example, a RF is less likely to overfit, when the bound for maximum tree depth is lowered. In the case of BLR, on the other hand, an option is to regularize the model by increasing prior regression weight variance, i.e. decreasing the hyperparameter α. Both of these maneuvres were tried on these models without notable success. Hence, the most likely explanation for the tendency to overfit is that there is too little data for the models to generalize. This problem could potentially be alleviated by increasing the proportion of training data to test data. This was also attempted, but without success due to the inability to draw reliable conclusions from such a small amount of test data.

As a proof of concept, these models manage to do what is required of them, though. Even with such little data, both models clearly outperform baseline predictions. As the most significant problems can be explained by the lack of training data, it can be deemed safe to assume that similar ML models would quickly improve with larger datasets.

5.2 Roadmap for data aggregation and storage

As was established in section 2.5, there is not much to build on, when it comes to data utilization in the Finnish construction industry. Hence, we have to begin building from the ground up, starting from the question of how to aggregate and store data; predictive modeling only becomes a viable tool when there is enough data to properly train and test the models.

With public entities already working on ways to promote extensive pre-demolition audits, and this study showing that predictive modeling is possible even with minimal demolition material mappings, it is critical to also think about how these documents should be stored in order for them to be easily accessible and usable. In an interview with the author, Senior Architect of The Finnish Ministry of the Environment Harri Hakaste stated that the ministry has already set in motion a project to design and build a public register for CDW data. Preliminary investigations for this register were completed in February 2022, and implementation is scheduled for 2023-24 (Personal communication, October 20, 2022). A feasible action point would be to have this register serve as a general repository, where contractors and real estate owners submit material mappings and pre-demolition audits.

Up to this point, we have knowingly disregarded a crucial aspect of our data. The PoC models in Chapter 3 were trained and tested on data based on pre-demolition estimates only. This is, of course, suboptimal. After all, the aim is to predict actual, realized material streams instead of estimates. Data on realized CDW amounts in demolition projects is currently even less available than pre-demolition audits. Therefore, it is crucial that the entities carrying out demolition projects submit detailed reports on actuals, as well as pre-demolition estimates. The previously mentionend public CDW data register would serve as a logical place to store those reports as well. Meanwhile, it is important to keep in mind that combining pre-demolition audits with actuals comes with the inherited challenge of scheme compatibility. Existing reporting models for actuals tend to be much more complex than pre-demolition audits, for example in that they usually differentiate CDW and other material streams with higher granularity using waste code driven schemes, for instance. Unaddressed, this will most likely lead to compatibility issues.

Considering the data even more closely, it is worth to also question what goes into the pre-demolition audits and actuals. Recalling that one of the most significant obstacles hindering CE in the construction industry is the lack of temporal alignment of demand and supply in the non-virgin material market, it is noteworthy that a temporal aspect in CDW data is currently absent in both pre- and post-demolition reporting. Hence, a key development point in CDW data aggregation is to start including estimates of temporal distributions for the released material streams in pre-demolition audits. Practically, this means stating for each material stream, at which time points during the demolition project a batch of that material would be available, in addition to how large those batches would be. Demolition projects can span several months, so it would be an oversimplification to assume that each material batch would be released as a single batch at the end of the project. Improving these estimates naturally requires that the realized temporal distributions are also properly reported in order to grow a base of quality data for training more sophisticated predictive models.

Compared to demolition data, the features, i.e. basic knowledge of target buildings, is much more available. For example, the National Land Survey of Finland maintains a public real estate register, from which anybody can access building data through REST APIs. On the other hand, due to a lack in documentation especially on the older side of the Finnish real estate base, the data is often incomplete. It is also crucial not to forget that the structures of buildings change during their lifetime due to e.g. renovations. According to the experts interviewed for this thesis, it often happens (less today than a few decades ago) that these changes are not properly documented and the data in the real estate register updated, which then leads to the risk of using incorrect data in the pre-demolition audit. Therefore, these changes have to be documented with the same care as during construction in order to preserve actual information on the building. All in all, it is crucial to uphold good quality standards in documentation and reporting during the whole building life cycle in order to ensure accurate data.

5.3 New opportunities

The CE transition tends to have the effect of inducing business cases, that in turn create opportunities for existing businesses and startups alike. In the case of the construction industry, this transition will create opportunities in the field for example for recycling operators, demolition contractors specialized in non-destructive demolition and recycling, and specialized retailers that, for example, acquire materials from demolition sites for refinement and reselling.

Assuming a different point of view, new possibilities present themselves in the area of data and analytics, as well. The platform project of Motiva Services (2.4) is a concrete example of how these possibilities can be utilized. Provided that data aggregation standards develop in the way suggested in 5.2, Motiva Services, in addition to other possible platforms and services, will have the opportunity to plug into different public registers and databases, combining demolition, real estate, and other data sources to create new insights and value. Possibilities include news and information sharing in the CE field, refining CDW data by combining it with e.g. material attributes and data about their environmental impact, modifying pre-demolition audits, post-demolition reports and other documents into directly usable data, and all the countless modeling and analytics services these data sources enable.

This potential also gives rise to the idea of the value of data. Without diving any deeper into the topic of data value, it is definitely an aspect the organizations governing these data should keep in mind. How this value is realized in the end, depends on the nature and purpose of these organizations. Possibilities include monetization strategies, for example, although these in turn depend strongly on the data sharing architectures implemented on top of the registers and data storages. The bottom line is, data should be treated as an asset that businesses are willing to invest into in order to refine it, and thus increase its value.

At the moment, most of the potential business value in predicting material streams is based on the cost cases for demolition contractors and real estate owners. Nevertheless, there are already early signs of developing business cases as well. Emerging technologies make it possible to recycle an increasing variety of CDW. For instance, the Danish company ReTec has been among the first to develop capabilities to recycle mineral wool. Moreover, steel production, being one of the most significant sources of GHG emissions globally, has been able to recycle scrap metal for some time already. The Swedish steel manufacturer SSAB states on their website that scrap manages to cover only 25-30% of the production's demand. With companies like SSAB striving to reduce their environmental impact drastically, it is safe to assume that a demand already exists for all scrap metal that is managed to salvage form CDW streams. And with rising demand for different materials found in CDW, waste handlers are able to place increasing prices on them.

5.4 Further research topics: what should be prioritized?

Due to the exploratory nature of this thesis, a significant part of its contribution is comprised of suggestions for further areas of research. The following topics are important points of focus that should be addressed next, as research in this field advances and more data is gathered.

5.4.1 Significance of the temporal distribution

Coming back to the alignment problem of supply and demand, the temporal distribution of CDW streams should be a priority research topic as the datasets available start to reach sufficient sizes. It could even be considered reasonable to postulate, that the ability to accurately predict batch sizes is considerabely less relevant compared to the readiness to give estimates of the points in time they are released at. The reasoning is that having even a rough knowledge of these time points beforehand makes it possible for entities on the demand side to collect the materials as soon as they are released, saving the demolition contractor from significant storage costs.

A challenge, on the other hand, is that often pre-demolition audits are carried out before demolition permits are given, which in turn means that the starting time of the demolition project is likely to be unknown during the audit. However, the temporal distributions can always be given relative to the starting time of the demolition project, that is, time points can be given as days after start, rather than strict dates. The dates can then be immediately calculated when the starting date is set.

5.4.2 Feature selection and sensitivity analysis

The question of feature significance was consciously left out of the scope of this thesis, mainly due to the lack of data. It is an extremely important area to explore further, though. The PoC models showed that predicting only the amounts of CDW should not require many features. This could potentially not be the case when predicting the temporal CDW distributions, or the amounts of reusable components. Therefore, aggregating accurate data of real estate during the whole lifecycle, as discussed in 5.2, is an important way of maintaining a variety of features to choose from and to experiment on.

As the amount of data grows, it becomes possible to perform sensitivity analysis on feature selection, finding out how changing the feature configurations affect the prediction of different targets. Models could start incorporating categorical features, like load bearing material or façade material, that are likely to improve prediction accuracy. Another point of interest is the question of the effect building type (office, storage, etc.) has on the model. A possible explanation for the scattered nature of the dataset used to train the PoC models is the fact that no distinction between building type was made. Taking the type into account as a categorical feature could help to cluster the dataset and increase prediction accuracy.

5.4.3 The reuse loop

This far, we have focused on predicting only an individual CDW stream, namely concrete, in order to potentially improve its recycling rate. As stated in 2.1, though, the potential impact of the reuse loop is much more significant than that of the recycle loop. It would therefore be sensible to shift more focus on trying to extract as many components as whole units before destructive demolition. Collaboration between pre-demolition auditors, recycling operators, and demolition contractors is key in working towards this, because reusable components are extracted both before and during demolition. Improving predictive capabilities in this area would support this work in giving recycling operators and demolition contractors some prior insight on what the target building contains, instead of them having to go in blind and providing the demand side with the information as the component extraction proceeds.

6 Conclusion

Finding ways to align supply and demand of non-virgin construction materials is an integral part of the transition towards a more circular economy that is happening in the construction industry. Being able to grow the window of opportunity to capture reusable and recyclable CDW as it is released, so that demolition contractors avoid storage or waste processing costs, is one of the key challenges to solve in this transition. It is a challenge that cannot be overcome without utilizing the full potential of data.

This thesis has shown that ML-based predictive modeling is a feasible approach to estimate CDW streams released during demolition - as long as there is enough quality data available. The PoC models have only scratched the surface of what is possible in terms of predictive modeling. Many critical problems, such as the temporal distribution of CDW streams, await solving as soon as the amount of available data is sufficient. Therefore, it is crucial that data in general - how to aggregate, store, share, and utilize it - is an aspect the construction industry starts including in all of its processes and capabilities. From the point of view of this thesis, it starts with standards for pre-demolition audit and demolition reporting, and the storage of those reports, but then extends across the whole industry. Every process generates data; it is up to the stakeholders to decide which pieces of information to capture and store.

Purely from a market economy point of view, the market for CDW and other non-virgin construction materials is only starting to develop. Highly profitable business cases are not yet driving the shift towards CE, but it is rather the existing and upcoming legislation and regulation concerning CDW handling, in addition to the potential cost savings that motivate e.g. demolition contractors to look for alternatives to storing or disposing of CDW. The focus of this thesis has been on the supply side, but in order to transition on a more holistic scale, the "other side of the loop", i.e. the demand, has to develop at the same time. Practically this implies a thorough paradigm shift in the construction industry, where all stakeholders in new construction have a common understanding of how to design and construct new buildings using as much non-virgin materials as possible. This shift, boosting the demand side, is likely to be a decisive force in inducing the business cases that will crucially accelerate the CE transition.

It remains an unanswered question, whether this shift will happen on its own, be it due to a critical mass of individuals and organizations taking the initiative to drive the transition, or some other phenomenon. It is, nevertheless, certain that during the course of the next several years, legislation and regulation on an EU and local level will develop a combination of "carrots and sticks" that will finally force the transition. In the end, it is up to every company and stakeholder in the industry to decide for themselves, how painful the transition will be. It can be deemed safe to say that failing to capture and utilize data to its full extent is likely to have dire consequences - while those, who have taken the initiative rather sooner than later to build their data competence and readiness, thrive.

References

[1] Matthew Adams, Victoria Burrows, and Stephen Richardson. *Bringing embodied carbon upfront - Coordinated action for the building and construction sector to tackle embodied carbon.* World Green Building Council, 2019.

[2] D. Barber. *Bayesian Reasoning and Machine Learning.* Cambridge University Press, 2012.

[3] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, October 2001.

[4] Mikko Dufva. *Megatrendit 2020.* Sitra, 2020.

[5] Circle Economy and ABN AMRO. *A futureproof built environment: Putting circular business models into practice.* ABN AMRO, 2017.

[6] Per-Anders Enkvist and Per Klevnäs. *The Circular Economy - a Powerful Force for Climate Mitigation.* Material Economics, 2018.

[7] Ellen MacArthur Foundation, McKinsey Center for Business, and Environment. *Growth within: a circular economy vision for a competitive Europe.* Ellen MacArthur Foundation, 2015.

[8] Andrew Gelman, Ben Goodrich, Jonah Gabry, and Aki Vehtari. R-squared for bayesian regression models. *The American Statistician*, 73(3):307–309, 2019.

[9] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[10] Annika Hedberg and Stefan Šipka. *The circular economy: Going digital.* European Policy Centre, 2020.

[11] Alexander Jung. *Machine Learning: The Basics.* Springer, Singapore, 2022. https://alexjungaalto.github.io/MLBasicsBook.pdf.

[12] Anna Laine, Tuomas Raivio, Håkan Jonsson, Anna Heino, Markus Klimscheffskij, and Jenny Lehtomäki. *Vähähiilinen rakennusteollisuus 2035 - Osa 1. Rakennetun ympäristön hiilielinkaaren nykytila.* Rakennusteollisuus RT and Gaia Consulting, 2020.

[13] Katja Lehtonen. *Purkutyöt – opas tekijöille ja teettäjille.* Finnish Ministry of the Environment, 2019.

[14] Kevin Murphy. Conjugate bayesian analysis of the gaussian distribution. November 2007.

[15] Bruno Oberle, Stefan Bringezu, Steve Hatfield-Dodds, Stefanie Hellweg, Heinz Schandl, and Jessica Clement. *Global Resources Outlook 2019: Natural Resources for the Future We Want.* United Nations Environment Programme, 2019.

[16] Ministry of the Environment, Ministry of Economic Affairs, and Employment in Finland. *New directions: The strategic programme to promote a circular economy.* Finnish Government, 2021.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[18] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, March 1986.

[19] Howard Raiffa and Robert Schlaifer. *Applied Statistical Decision Theory.* Harvard Business School Publications. Division of Research, Graduate School of Business Adminitration, Harvard University, 1961.

[20] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.

[21] X. Solé, A. Ramisa, and Carme Torras. Evaluation of random forests on large-scale classification problems using a bag-of-visual-words representation. *Frontiers in Artificial Intelligence and Applications*, 269:273–276, January 2014.

[22] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010.

[23] McKinsey & Company World Economic Forum, Ellen MacArthur Foundation. *Towards the Circular Economy: Accelerating the scale-up across global supply chains.* World Economic Forum, 2014.

[24] Ying Zhu, Harriet Lonka, Katja Tähtinen, Markku Anttonen, Päivi Isokääntä, Anssi Knuutila, Jukka Lahdensivu, Selma Mahiout, Anne-Marie Mäntylä, Markku Raimovaara, Tiina Rantio, Tiina Santonen, and Tuuli Teittinen. *Suitability of demolition materials for different applications from the safety and health point of view.* Finnish Prime Minister's Office, 2022.

[25] Ying Zhu and Katja Tähtinen. *Rakennusosien uudelleenkäytön edellytykset Suomessa.* Finnish Government's analysis, assessment and research activities, 2022.