Quality prediction of the agglomeration line product in a plastic waste separation plant

A contribution towards machine setting optimization in waste separation plants

Author: Baldur Otto
Date: 25.10.24

TU Delft

Quality prediction of the agglomeration line product in a plastic waste separation plant

A contribution towards machine setting optimization in waste separation plants

By

Baldur Otto

In partial fulfilment of the requirements for the degree of:

Master of Science
in Environmental Engineering

at the Delft University of Technology,
to be defended publicly on Friday November 15, 2024 at 12:30 AM.

Supervisor:	Dr. Hongrui Wang	TU Delft
Thesis committee:	Dr. Hongrui Wang	TU Delft, chair
	Dr. Abraham Gebremariam	TU Delft
	Dr. Jonas Rzezonka	Sutco RecyclingTechnik GmbH

Contents

List of Figures

List of Tables

Acknowledgments

I am profoundly grateful to Hongrui Wang and Abraham Gebremariam for their supervision and unwavering support throughout the writing of this thesis. Particularly the frequency of the guidance, the dedication and flexibility are very special coming from the context of German Universities and have been pivotal in my academic development and the organization of this endeavour.

Special thanks are extended to Jonas Rzezonka and Dorothee Sänger for their meticulous oversight and invaluable contributions to this thesis. Their insightful inputs have not only facilitated the achievement of this work but have also significantly enhanced the quality of this thesis.

I am particularly appreciative of the holistic approach taken by all supervisors, which encompassed not only academic outcomes but also considerations about feasibility of tasks with regards to temporal scopes, frequent checks if I am equipped with all necessities to conduct my research and a focus on work-life balance as well as mental health. This comprehensive support has not only resulted in a thesis of which I am proud but has also made the entire process enjoyable and rewarding. While many of my peers experienced thesis writing as a stressful period, I found it to be a time of personal growth and fulfilment. Thank you!

Abstract

Through improved plastic waste separation EU recycling goals can be reached and environmental economic advantages can be unlocked. To help with this endeavour, this research explores dynamic separation efficiency determination and waste stream characterization through near infrared (NIR) separation unit and belt weigher data in a plastic waste sorting plant in Scandinavia. For the showcasing of these concepts, the goal was to predict the product quality of the high-quality (HQ) agglomeration line, using data of the first NIR-scanner in the agglomeration line as prediction input. In the agglomeration line two NIR-scanners are connected in series to ensure high-quality separation of the material. Through the NIR-scanners, material-specific area flow data is available and through the belt weighers mass flow input to each NIR-scanner is provided. Quality criteria are weight shares of PO (target material) and PVC (main contaminant). Difficulties arose, as the material-specific mass flow is needed for quality determination but only the total mass flow is provided. This was addressed by modelling area densities using a linear regression model, with belt weigher and NIR-scanner data as input. Using the calculated area densities, the material-specific mass flow was determined. For validation, summed material flows were compared with belt weigher data, yielding a mean absolute error (MAE) of 141 [kg/h] and a mean relative error (MRE) of 3%. The separation efficiency was determined through an XGBoost model, to predict material-specific area flow of the second NIR-scanner. Results were a MAE of 50.02 [m²/h] and an MRE of 1.1% for the total area flow. The final separation step could not be validated, as no NIR-analyser is present behind the second NIR-scanner. Therefore, separation efficiencies from the previous separator were transferred. Joining all three concepts the weight share of PO and PVC could be predicted with a MAE of 0.36% and 0.007%. For the joint outcome, greater uncertainty contribution was ascertained for the area densities compared to the XGBoost application. Future research is recommended for separation efficiency determination of the last separation step and for improved modelling of the area densities.

1 Introduction

1.1 General Context

In 2018, EU recycling goals got updated and extended to include packaging material through the amendment of the waste framework directive. This results in a minimum recycling rate of 65% for municipal waste and a maximum 10% of waste that is landfilled by 2035. Furthermore, until 2030, 70% of all packaging waste has to be recycled. Here, the subcategory of plastics must reach a minimum recycling rate of 55%. (European Commission, 2018)

Currently only Slovenia and Germany achieve the goal for municipal waste. On the worst performing end Romania has a recycling rate below 15%. Apart from that only 8 countries reach the goal for packaging waste where Romania is again the worst performing country with a recycling rate below 40%. (Eurostat, 2024a; Eurostat, 2024b)

On a broader scale, global waste generation is predicted to rise from 2.01 billion tons per year in 2016 to 2.59 billion tons in 2030 and to 3.4 billion tons in 2050. Connecting the produced waste to its environmental damage, 1.6 billion tons of CO_2 equivalents were connected to solid waste treatment in 2016. This resembles 5% of the global greenhouse gas emissions. Translating this value to the predicted waste generation in 2050, this indicates 2.6 billion tons of CO_2 eq. emissions for 2050. Zooming into specific regions, for Europe and central Asia, a growth from 393 million tons in 2016 to 490 million tons in 2050 is predicted. (Kaza et al., 2018)

These Figures show the need for an effective and sustainable waste management. This is necessary to treat the increasing waste production, to reach the recycling goals of the European Commission and to leaver the potential environmental benefits through recycling.

Putting numbers on these environmental benefits, life cycle analysis (LCA) methodology is used. To do this, a focus on packaging materials is set. Maga et al. (2019) found that recycled PET in food packaging can reduce the footprint of the packaging by 40%. Tonini et al. (2021) concluded that recycled HDPE only has 67% of the emissions of virgin HDPE. Recycled PP only showed 44% of the emissions of virgin PP. Lastly, Civancik-Uslu et al. (2019) determined that a cosmetic bottle with mineral fillers and recycled HDPE can save more than 30% CO_2 eq. emissions compared to a virgin material bottle.

Transferring these benefits to a more systemic level, Schwarz et al. (2021) determined that with recycling of the 15 most used polymers in Europe CO_2 eq. emissions can be reduced by 73%. This means that 200 million tons of CO_2 eq. can be avoided through the recycling of these plastics. This figure was obtained for a scenario with improved waste sorting. Improved waste sorting is mainly hindered by impurities, but shows the greatest enhancement of environmental benefits. Dokl et al. (2024) show similar findings. They acknowledge advances in waste processing, but claim that too much material ends up in the mixed plastic fraction due to insufficient sorting. This fraction faces downcycling and therefore symbolizes unused potentials during plastic waste sorting.

Unused potentials during material separation and recycling are furthermore reflected by the economic risks of a waste separation plant. This is due to the fact that not only products of positive value but also

of negative value are produced (Feil et al., 2017; Ozdemir et al., 2021). Therefore, it becomes not only evident from an ecological perspective, but also from an economic perspective, that improved sorting and recycling is beneficial.

A waste separation process is commonly composed of an initial comminution and classification step using shredders, drum sieves, wind sifters and ballistic separators (Feil et al. 2017; Ozdemir et al., 2021). During or after this step, metals get removed by over belt magnets and Eddy current separators (Ozdemir et al., 2021). Lastly, the most crucial units for plastic separation are sensor-based sorters. These units separate remaining waste streams into mono-material streams, which enables their further processing. (Friedrich et al., 2022)

One of the hindrances for improved waste separation processes is that waste separation plants are only evaluated very sparsely. Very short time frames of several weeks for the entire life span of a plant are reported. Longer periods of 1-2 years are identified in rare cases and are commonly connected to scientific projects (Gadaleta et al., 2020). The consequence of this is that most plants only optimized their machine settings and process parameters once. This is done during the commissioning of the plant. Nevertheless, the composition of waste is constantly changing (even by season), wherefore this optimization gets outdated rather quickly. Accordingly, a more frequent or even real-time optimization of process parameters is needed to extract the maximum amount of secondary raw materials. (Kroel et al., 2024a)

To address the described problems and to leverage the shown advantages of improved material separation, a more frequent material sorting plant optimization needs to be implemented. This study aims to help with this endeavour by exploring opportunities of in-plant recorded data use, with the ultimate goal of real-time plant optimization. As a first step towards this objective, the purity prediction of the agglomeration line product in a plastic waste sorting plant in Scandinavia will be showcased.

1.2 Company Cooperation

The thesis research project is conducted with Sutco Recyclingtechnik GmbH located in Bergisch Gladbach, Germany. Sutco is one of the leading waste separation plant planners for large scale waste separation facilities, with worldwide construction activities. Their latest projects were carried out in Poland, Chile and Austria, where they constructed a plastic sorting facility that has the capacity to process 50% of Austria's plastic packaging waste. (Sutco, 2024a; WMW, 2024)

Recently, Sutco increased effort for full digitalization of waste separation plants, with the ultimate goal of enabling real-time plant optimization and to set up a digital twin for each facility (Sutco, 2024b). This work hopes to deliver a useful building block to this goal.

Apart from this, the thesis research project is part of the "Energieeffiziente Sortieranlage" (EnSort) project. The project has the objective to facilitate waste sorting plant optimization in real-time and to foster energy efficiency. Next to Sutco, other partners are TU Dresden, Universität Bremen and Ruhr-Universität Bochum as well as TOMRA Sorting GmbH (TU Dresden, 2023). TOMRA is the company that produced the near infrared (NIR) scanners for the analysed waste sorting plant in Scandinavia.

1.2.1 Waste Separation Plant

The data for this thesis research project is retrieved from a plastic sorting plant in Scandinavia, which was planned and built by Sutco. Belt weigher data is directly available from Sutco, while NIR-scanner data is provided by TOMRA. In Figure 1, a simplified process flow chart of the plastic sorting plant be found. Preconditioning and classification of the material are conducted upfront and have been left out for simplicity.

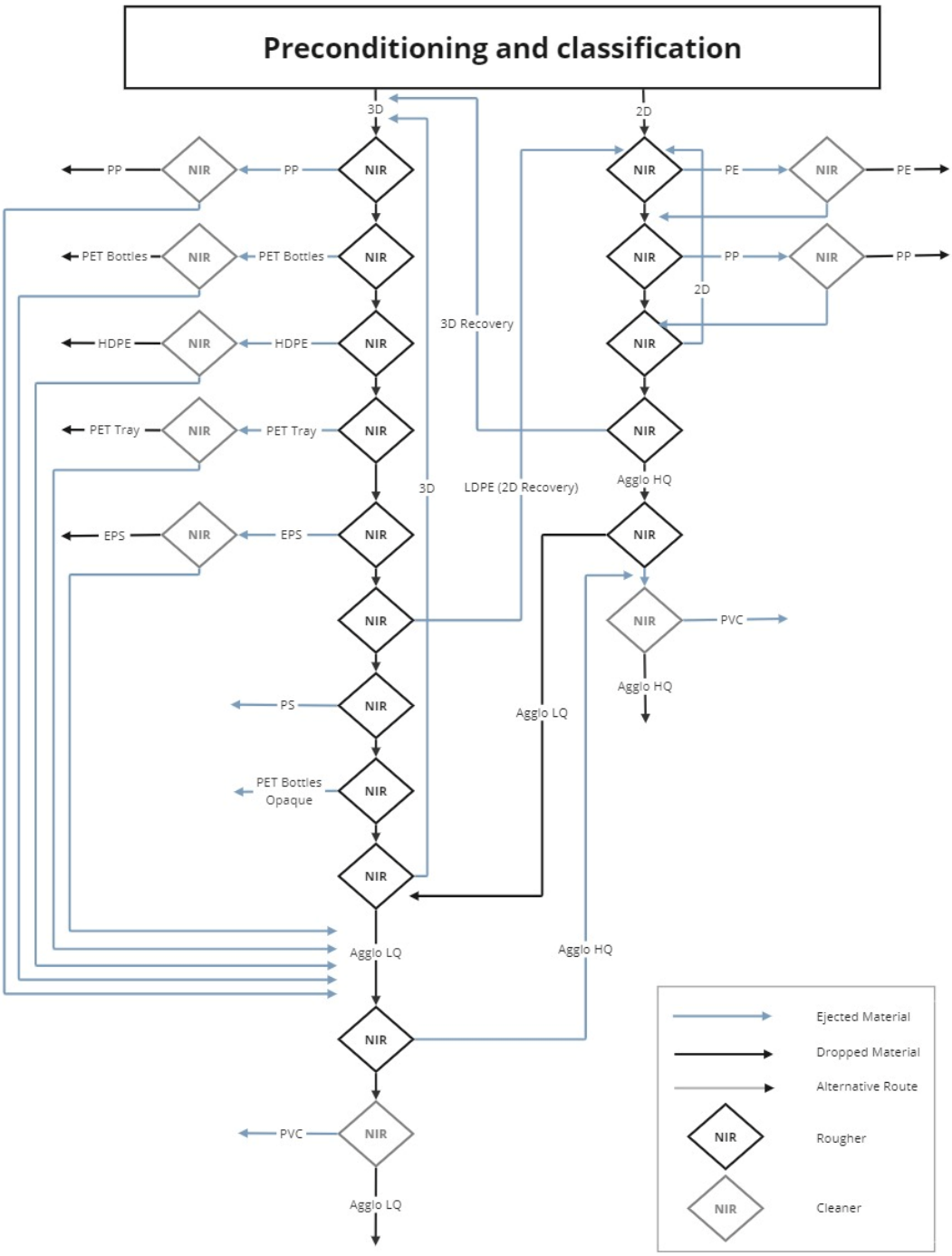

Figure 1: Process overview of the analysed plastic waste sorting facility.

After the preconditioning and classification, the material is split up into a 2D and 3D section. For the 2D section subsequently PE, PP and misclassified 3D objects are removed and cleaned. The residues end up in the agglomeration ("Agglo") section, which resembles a mixed plastic stream. The higher quality agglomeration stream can be turned into plastic products, which have low requirements regarding material characteristics. Low-quality agglomeration streams end up in incineration, but are wanted due to their high burning value. Due to the heat application in both cases, PVC gets removed from the streams to avoid the formation of chloric acid. For the 3D section PP, PET bottles, HDPE, PET trays, EPS, misclassified 2D materials, PS and opaque PET bottles get removed and cleaned. The residue of the 3D section is the basis for the low-quality agglomeration stream.

1.2.2 Agglomeration – Working principles

To limit the data, available to this study, to a reasonable and workable amount, it is further focused on the agglomeration section of the plant. This section was chosen due to its exceptionally high data availability compared to other separation lines. This advantage comes with the payoff that this part of the plant is fed with the residuals of the remaining separation lines. Therefore, the received material is more likely to act unexpectedly but a higher data resolution is obtained.

To generate the high-quality (HQ) and low-quality (LQ) agglomeration stream, six NIR-scanners and two belt weighers are involved. After the sorting steps, another process for compaction of the HQ product is added. Moreover, four additional belt weighers are installed for quality control and data acquisition. An overview of the agglomeration line can be found in Figure 2.

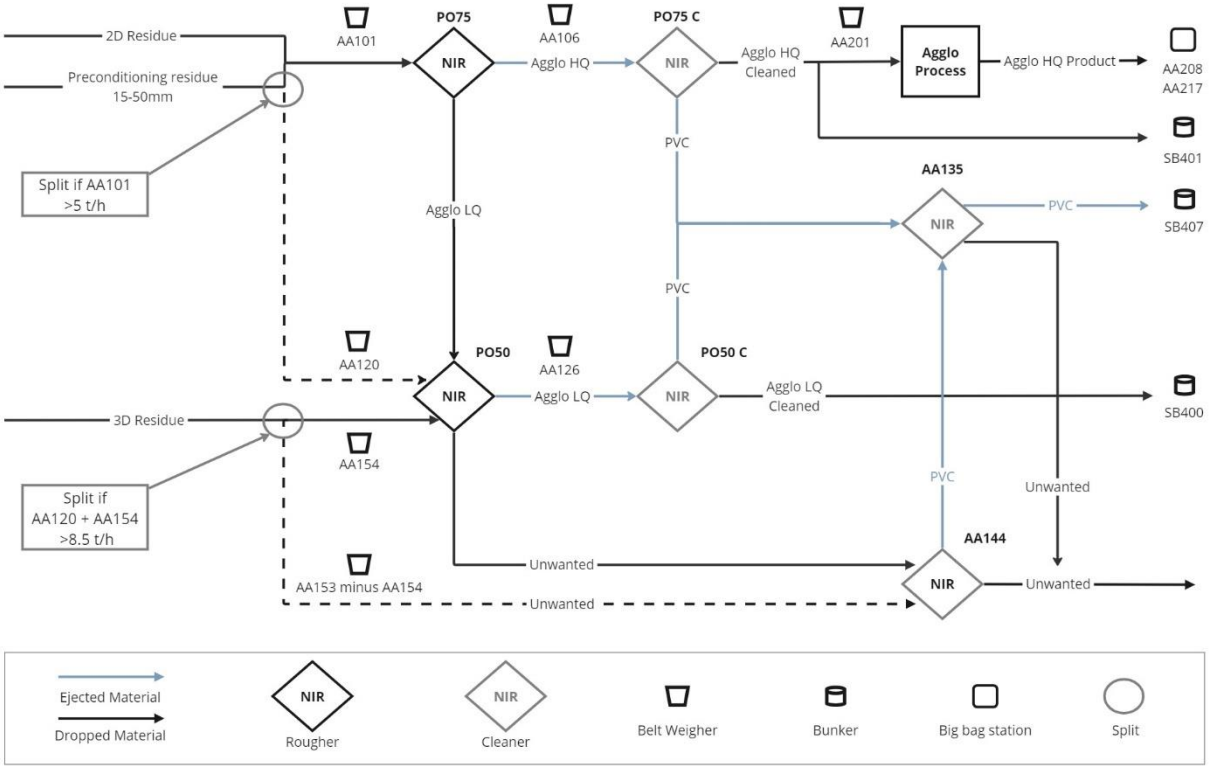

Figure 2: Agglomeration line of the plastic waste sorting plant.

The HQ agglomeration line is fed with 2D residues and residues from preconditioning and classification with sizes between 15-50mm. To avoid overloading of the HQ agglomeration line belt weigher AA101 was installed. Therefore, if material streams greater 5 [t/h] are detected, the preconditioning residue is led to the LQ agglomeration line for processing. This is done as the preconditioning residue contains less target material. Target materials are PE and PP, which are together known as polyolefins (PO) in this context. On PO75, PO is ejected, while all other materials are dropped and transported to PO50 for processing in the LQ agglomeration line. The names of the NIR-scanners indicate the desired PO contents. Therefore, for the HQ agglomeration line a PO content of 75% is wished and for the LQ agglomeration product 50% PO content is aimed for. In a next step a cleaner ensures ejection of PVC to avoid formation of chloric acid during further processing. Finally, the HQ stream is sent to a bunker or to an additional compaction and cleaning step.

The LQ agglomeration line receives input from the 3D residues as well as the preconditioning in the case that the HQ line is overloaded. With 8.5 [t/h] the LQ line has higher capacity than the HQ line and can redirect input to the unwanted fraction in case of overburdening. Like in the HQ agglomeration a hierarchy is applied, where the preconditioning residue is preferred over the 3D residue. This is done as the preconditioning residue has a higher PO content compared to the 3D residue. On PO50, PO is ejected and passed on to PO50C, while the dropped material enters the unwanted category. This unwanted material is sent to AA144, where PVC is ejected. The drop of AA144 resembles the residue of the sorting plant. On PO50C, separated PO is cleaned through ejection of PVC and sent to the LQ agglomeration bunker.

After separation of the HQ and LQ agglomeration product, PVC ejects from PO75C, PO50C and AA144 undergo another PVC separation on AA135. This showcases the importance of PVC removal and gives insight that the separation efficiency of NIR-scanners is not perfect. After processing on AA135, PVC is sent to a bunker, while the drop of the scanner enters the residual stream of the waste sorting plant. To fully understand the quality requirements of the HQ and LQ agglomeration products, desired product compositions are listed in Table 1. These compositions were agreed up on between Sutco and the sorting plant proprietor.

Table 1: Quality requirements of the HQ and LQ agglomeration line product.

Quality	PE and PP	PVC	PET	Other Plastics	Organics	PS
HQ	>75%					
		<1.4%				
			<6%*			
				<13%*		
					<10%*	
LQ						
	>50%					
		<2%				
			<15%*			
				<20%*		
					<20%*	
						<5%*

*Agreed upon but not tested/evaluated

Inspecting the quality requirements, especially the low PVC thresholds give reason to the focus on PVC removal during the agglomeration line processing. This is due to the fact that the agglomeration line product is either used for incineration or will be heated up for formation of new products. During these processes chloric acid could form, which explains the low threshold for PVC. With the high PO contents that are needed the targeting of PO on the initial NIR-scanners of both lines becomes compulsory. High PO content raises burning values and enables processing as a plastic, wherefore the material is wanted. Apart from that, it is interesting that fewer material contents that were agreed on were actually tested. This is due to cost constraints of probing and can be reasoned by a reduced amount of quality requirements, of the agglomeration product buyers, then initially expected.

1.2.3 Area under study

Zooming further in, the area of application for this study was determined. The goal of this research is to deliver an exploration regarding product purity prediction in waste separation plants. This is intended as starting point to enable real-time machine setting optimization during waste separation. No study with a similar goal could be found in literature research. Therefore, as a first step, the smallest possible unit for this undertaking was sought after. Accordingly, a part of the HQ agglomeration line was selected, as it has the highest data availability paired with the highest expected data quality. This expectation is due to a reduced amount of contaminations in the HQ agglomeration input compared to the LQ agglomeration input. The described part of the HQ agglomeration line can be observed in Figure 3.

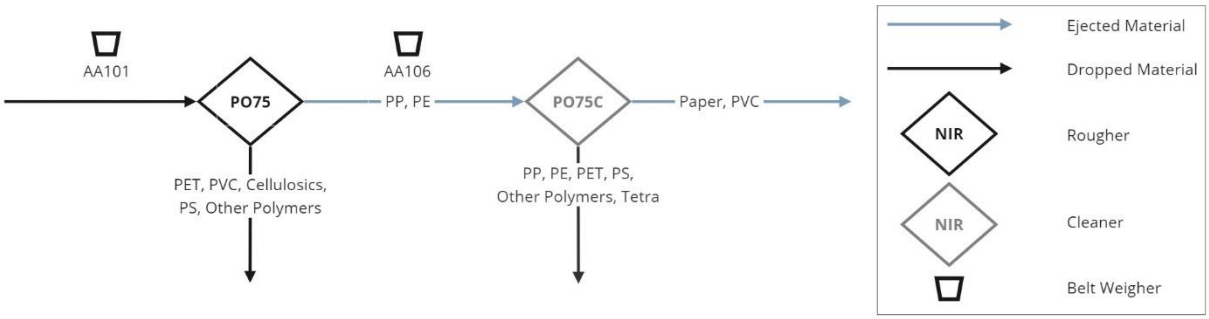

Figure 3: Area of application of this study, placed in the HQ agglomeration line and consisting out of two NIR-scanners and two belt weighers.

Compared to Figure 2, belt weigher AA201 and bunker SB401 are left out. This was done as the material leaving PO75C is either led into the agglomeration process and weighed by AA201 or fed into the bunker and baled at a later stage. As the bunker filling level is detected in m^3 and the density of the material is unknown, uncertainty for the use of this data was considered too high.

2 State of the art

To lay the foundation for the identification of the knowledge gap and the development of the research question, important concepts to the problem and goal of this research will be explained. Regarding modelling, focus was laid on methodologies that turned out crucial during the execution of this work.

2.1 NIR-scanner working principle

NIR-scanners have become the prevalent technology and a global trend for the separation and identification of plastic waste (Zheng et al., 2018, Dokl et al., 2024). NIR separation units deploy material classification, using the near infrared spectrum between wavelengths of 750 nm and 2500 nm. To do this material is exposed to NIR radiation and the reflected spectra gets analysed. These spectra differ per type of plastic due to their unique chemical composition. Especially, the main groups of the polymers like carbon (C) – hydrogen (H), nitrogen (N) – hydrogen (H) and oxygen (O) – hydrogen (H) groups as well as other hydrogen containing groups react differently to the NIR spectra. These effects can be due to absorption of energy, overtones, interferences and more. Due to this, each material emits unique reflections, which resemble the classification mechanisms of NIR separation units (Zheng et al., 2018; Du et al., 2022; Dokl et al., 2024).

Although NIR-scanners are one of the technologies with the highest signal to noise ratio, the identification of materials is not an easy task (Zheng et al., 2018). To successfully classify and sort the particles several process steps need to happen. At first, objects have to be identified from the obtained image. This means that for each pixel of the image it has to be decided if it belongs to an object or not. Through these decisions, objects can be represented as clusters of pixels. After this, each pixel of an objects gets assigned a material class and with the help of customized recipes the object can be categorized. After successful detection, objects of interest gets ejected from the stream via air nozzles (Friedrich et al., 2022)

Separation can either happen via positive or via negative sorting. In positive sorting, materials targeted by the separation process get ejected, while for negative sorting impurities are removed. In the first case the NIR separation unit can also be referred to as rougher, whereas for the second case they are called cleaners. (Kroell et al., 2024a)

In Figure 4, common sensor-based separation unit setups can be observed. Particles can be either feed via a conveyor belt, a chute or by falling into the separation unit. Air nozzles are used to enlarge or decrease the trajectory of particles ejected from the conveyor belt (a,b,d) or to shoot out particles that are falling vertically (c). (Maier et al., 2020)

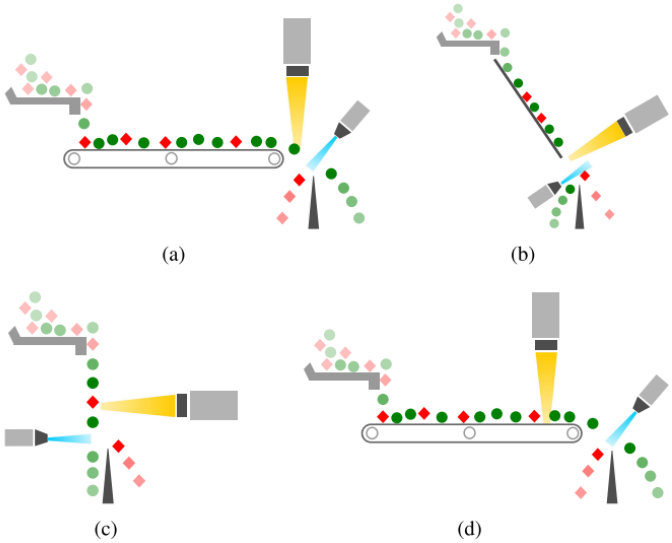

Figure 4: Different setups for a sensor-based separation unit with indication for radiation used during detection (yellow shapes), air flows for ejection (blue shapes), targeted particles (red shapes) and non-targeted particles (green shapes) – a) Feeding via conveyor belt, b) Feeding via chute, c) Free fall feeding, d) Feeding via conveyor belt. (Maier et al., 2020)

2.2 Machine Learning in Waste separation

Machine learning (ML) is used and researched for several improvements of the waste management process. An overview of applications found during literature research is presented in the following.

Several researchers propose the use of smart bins for improved source separation of waste. Desai et al. (2018) investigate the distinction between degradable and non-degradable waste via a camera unit together with a convolutional neural network (CNN). Implementation is planned using a raspberry pi unit and a robotic arm to take over separation for the user. Sheng et al. (2020) propose a similar but more sophisticated approach. Their goal is to separate waste into metal, plastic, paper and residual waste fractions. For this they apply a tensor flow deep learning model to recorded pictures and connect their outcomes to an internet of thing (IoT) approach. This approach is characterized by tracking the fill height of the bins to optimize collection. Rahman et al. (2022) implement a very similar methodology but show the potentials of the framework by implementing a CNN model instead of a tensor flow model. Chen et al. (2022) apply the same approach but take it a step further by implementing an additional ML model for the optimization of the waste collection routes.

Identification of material with NIR data can be challenging due to noise present in the received spectra. To improve this several approaches implementing ML have been found. Du et al. (2022) established a CNN for the detection of different textile materials. For this, they converted NIR spectra of each object into a 40x40 greyscale image making. With this they made it usable for a CNN model and could achieve an accuracy of over 95%. The added benefit of the ML approach is the improved interpretation of the NIR spectra, which until now is one of the bottlenecks for successful textile waste classification. Zheng et al. (2018) used the combination of a hyperspectral imaging system (HIS), a NIR-scanner and principal component analysis (PCA). With this, they achieved 100% classification accuracy for acrylonitrile butadiene styrene (ABS), polystyrene (PS), polypropylene (PP), polyethylene (PE), polyethylene terephthalate (PET), and polyvinyl chloride (PVC). This is a success, as with conventional

NIR-scanners there is a time lag between provision of location and spectral data, while the NIR-HIS combination delivers this information in one step. Another difficulty is the detection of black objects. For this, Dokl et al. (2024) propose the joint use of cameras and NIR-sensors together with ML to achieve sufficient separation. Kroell et al. (2024a) shift the focus from classification of material to the separation efficiency of the NIR-units. For this, they trained a polynomial regression (PR), a random forest (RF) and an artificial neural network (ANN). Their main feature for the model training was the conveyor belt occupation density [m^2/m^2] and the material-specific conveyor belt occupation density [m^2/m^2]. All three models predicted the separation efficiency of the unit with a mean absolute error (MAE) below 6.3%. The ANN performed best with a MAE of 3%. These findings represent a step towards real-time machine setting optimizations and improved plant simulation. Nevertheless, their generalizability towards application in real world separation facilities is in question.

Zooming out from material classification and separation efficiency determination, Kroell et al. (2021) contribute to sensor-based material flow characterization and process monitoring. As NIR-scanner information is finally broken down into 2D data, information about volume or weight through corresponding densities is not available. Furthermore, weighing is inhibited during the process, as this would mean a stopping of the conveyor belt if no belt weighers are present. Therefore, Kroell et al. (2021) used a setup consisting of laser triangulation and an RGB camera to retrieve the missing information. Training a RF model with RGB camera and laser triangulation data, R^2 values of 0.76 could be reached. This resembles a major improvement compared to conventional area density databases, which reached a maximum R2 of 0.53 for the tested material. In a follow up research, Kroell et al. (2024b) managed to determine particle weights of plastic pre-concentrates in a lightweight-packaging sorting plant with NIR sensor data only. In the described case, they managed to outperform manual quality control with measurement uncertainties of 0.31 w-%.

Searching for ML applications on a more systemic level, Xi et al. (2022) use an ANN model to predict waste processing capabilities in Chinese cities. For this, they use features like population, education level, number of waste collection vehicles, budget of the local government and many more. With this, a prediction accuracy of 95% could be achieved. Furthermore, the results were used to extract the most important feature of the model. Accordingly, the public budget expenditure was the most influential variable with a weight of 52%. Furthermore, it became evident that with an appropriate separation, municipal waste can be reduced by 30-40% in the observed cases. This shows the capability of ML methodologies to analyse large-scale systems, to pinpoint hotspots and to identify areas of interest in the field of waste management.

2.3 System characterization and monitoring

To understand how the quality of the agglomeration line product can be predicted, an introduction to waste sorting plant monitoring and waste stream characterization is needed.

Prevalent methods for waste separation plant characterization are mass balance approaches like material flow analysis (MFA). These can be implemented using common MFA software like STAN. To do this, processes and transfer coefficients (TC) are defined. TCs symbolize the distribution of material inputs to the outputs of the process. This is done in percentual fashion and material wise. (Feil et al., 2017; Gadaleta et al., 2020; Tanguay-Rioux et al., 2022)

To successfully determine a system, all inputs regarding amount and material composition as well as TCs of all processes have to be known. Alternatively, enough in-plant streams, their composition and a sufficient amount of TCs must be determined. Furthermore, an underdetermined system can become determined through an optimization problem approach. This methodology is proposed by Tanguay-Rioux et al. (2022) and applies a mass conversion approach together with a defined set of constraints and sampling.

For TC determination, predominantly expert knowledge is applied. Here, the problem is that already known TCs can hardly be transferred to a new separation problem. This is due to that waste stream composition is changing based on geographical scope. Therefore, TCs differ from separation process to separation process. Tanguay-Rioux et al. (2021) try to summarize common TCs for separation units and to establish a minimum and maximum separation efficiency range. An example for NIR-scanner separators can be found in Table 2.

Table 2: Minimum and maximum separation efficiencies for NIR separation units (Tanguay-Rioux et al., 2021).

Unit	Minimum TC	Maximum TC
NIR-PET	0.83	0.93
NIR-HDPE	0.71	0.83
NIR-Mixed Plastics	0.74	0.74

Another approach to characterize a waste separation system is the sampling of important flows within it. Following the methodology of Tanguay-Rioux et al. (2022), composition and quantity of a flow are sampled. Alternatively, one of both information is retrieved, if this is sufficient to determine the system. Another framework for sampling is presented by Feil et al. (2017), where sampling can be classified into 3 main types. The first approach is to determine the quantity and the quality of the product. The second approach is to sample the residue to get to know the amount and type of wrongly sorted particles. Lastly, the third approach is to sample the process with the goal of complete data acquisition at all process steps.

Nevertheless, the characterization of a system with static TCs and sampling comes with limitations. As described by Kroell et al. (2024a), composition of waste streams changes significantly over time, sometimes even by season. Furthermore, information on waste stream composition is barely available, as showcased by a study conducted by Eriksen & Astrup (2019). In their research, they sampled 3700 kg of source separated plastic waste, to get to know its composition. This was necessary, as no information on source separated plastic waste was available upfront.

Focusing on the material flows inside separation plants, Curtis et al. (2021) report influences on product purity due to material flow changes during operation. These fluctuations can be caused by braid formations, bridging and material flow delays due to object shapes or machine specific discontinuous material discharge. An example for the object specific delay could be a round object that rotates on top of a shredder unit and therewith delays its processing. Regarding the machine specific delay, drum sieves are a good example where smaller objects get discharged faster than bigger objects.

Trying to establish dynamic TCs, Küppers et al. (2020) investigated the influence of material composition and conveyor belt occupation on the separation efficiency of NIR separation units. For the material composition, only a negligible influence could be found. Nevertheless, for the belt occupation, a 4^{th} degree polynomial could be fitted. Therewith, TCs have been adjusted according to the conveyor

belt occupation and reached a R^2 of 0.94. Kroell et al. (2024a) took a similar approach using an ANN model and could predict separation efficiency with a MAE of 3%.

To implement dynamic TCs and process monitoring, data from sensor-based material analysers is needed. Nevertheless, sensor-based analysers are connected to high economical cost, wherefore intensive process monitoring with this technology is economically unfeasible. A potential solution could be the use of already existing units within the plant. Conveniently NIR separation units have to detect material composition to enable separation. As the data collected by the sensors is task specific, it requires additional process step to make it usable but resembles enormous process monitoring potentials (Kroell et al., 2022a). For complete process monitoring, additionally the time that the material travels from one unit to the next has to be known. Furthermore, fluctuations and delays of material flows must be detected and considered in monitoring and process modelling (Curtis et al., 2021). Apart from that, an out of the box benefit of extensive process monitoring is that dangerous objects can be identified and removed before they cause harm. (Vrancken et al., 2017)

A last step to complete waste separation process characterization is the evaluation of the process performance. Three waste separation plant performance indicators were introduced by Feil et al. (2016) and used by the cited studies of Küppers et al. (2020), Curtis et al. (2021) and Kroell et al. (2022b, 2024a). The indicators are purity, recovery and yield. Compared to this, Tanguay-Rioux et al. (2022) and Gadaleta et al. (2020) make use of two main indicators, namely recovery and purity. As their definition of recovery is equivalent to the definition of yield from Feil et al. (2016), the definitions of Feil et al. (2016) will be used for the rest of this work to avoid confusion.
The recovery R [-] describes the ratio of the sorting process input with regards to the sorting process product. Here, the sorting process product means the flow that can be sold with positive economic value. In the formula below m is indicated as mass:

$$R = \frac{m_{product}}{m_{input}} \tag{2.1}$$

The yield Y [-] refers to the amount of targeted material that ended up in the product. This is evaluated based on the amount of targeted material that was present in the input stream. The formula can be found below, where m is referred to as mass and c is referred to as concentration:

$$Y = \frac{m_{product} * c_{target}}{m_{input} * c_{target}} \tag{2.2}$$

The purity P [-] symbolizes the ratio of target material in the product. As no separation process is perfect, it can be seen as the main quality criteria. In the following formula m is represented as mass:

$$P = \frac{m_{target}}{m_{product}} \tag{2.3}$$

2.4 Relevant statistical models and concepts for this work

Moving on from the more general introduction of relevant research and concepts, in this subchapter important statistical methodologies, phenomena and models for this work will be introduced.

2.4.1 Ordinary Least Square regression

Starting off, the ordinary least square (OLS) regression plays a big role in the later deployed area density determination based on NIR-scanner and belt weigher data. When the OLS method was invented, the novelty of it was that the residuals between the estimated and measured values were considered (Dismuke & Lindrooth, 2006). This leads to a formulation of the regression problem, where $Y\ (n,1)$ represents the dependent variable, $X\ (n,p)$ the independent variable, $\beta\ (p,1)$ the estimators and $\varepsilon\ (n,1)$ the residuals. Apart from that, n indicates the number of observations and p the number of independent variables:

$$Y = X * \beta + \varepsilon \tag{2.4}$$

The estimators are determined by summing up the squared residuals and minimizing them. For this, the estimate of the dependent variable gets calculated with the estimators as well as the independent variables. The result gets then subtracted from the measured values of the dependent variable to determine the residual. The advantage of this technique is that the minimization of the squared residuals can be solved through a mathematical formulation and does not require any iterative procedures. The respective formula can be found below (Dempster et al., 1977; Dismuke & Lindrooth, 2006):

$$\beta = (X^T X)^{-1} X^T Y \tag{2.5}$$

To obtain a valid result, several assumptions have to be fulfilled. These are the normal distribution of the residuals, homoscedasticity and independency of the independent variables. Homoscedasticity can be described as a constant variance of the residuals across the observed data. Furthermore, autocorrelation should be controlled for. (De Souza & Junqueira, 2005; Dismuke & Lindrooth, 2006)

2.4.2 Multicollinearity

A problem that can occur during the application of OLS is multicollinearity. Multicollinearity means that the independent variables are highly correlated. In general, this does not violate regression assumptions, but can cause problems with the interpretation of the regression coefficients (Gujarati, 2021). This means that the prediction of the dependent variable by the independent variables is not hampered but that e.g. confidence intervals of regression coefficients can contain zero. Confidence intervals with this characteristic are problematic, as it is unclear if an increase of the respective independent variable leads to an increase or decrease of the goal variable (Paul, 2006). This can also be explained in a way that the independent variables are correlated in a manner where they share very equal explanatory power. As a result, the model splits up the influence on the goal variable randomly. Lastly, multicollinearity can lead to sensitivity for small changes in data due to the numerical workings of the algorithm (Gujarati, 2021).

2.4.3 Bootstrapping

To obtain greater confidence in OLS modelling outcomes bootstrapping can be applied. Bootstrapping is used to generate inferences about population parameters with a limited number of samples. In other words, the goal is to quantify uncertainty about population parameters at hand. To achieve this, the data is resampled several times with the same amount of data points of the initial sample. The

resampling is done through random extraction of data points from the original dataset where several drawings of the same datapoint are possible. Assuming that the original sample is representative, bootstrapping can therefore help to estimate population parameters and their uncertainty. This is due to the variation in the generated samples and the parameters that can be calculated for each of them. (Choi, 2016; Youness et al., 2023)

Another advantage is the application of further statistical methods like confidence intervals which can be explored through bootstrapping by a data driven approach. This can help to generate greater confidence in the retrieved population parameters, as they have not only been inferred from the initial sample but have been retrieved from a number of subsamples. (Choi, 2016; Mokhtar et al., 2023)

2.5 Relevant machine learning models for this work

For the state-of-the-art description of machine learning models, it was focused on four models that became important to the thesis project during the conduction of this work. These models are Ridge Regression, Gradient Boosting, Extreme Gradient Boosting (XGBoost) and Multilayer Perceptron (MLP). In the following subchapters the inner workings of each model are explained and other models that were used are briefly introduced.

2.5.1 Ride Regression

Ridge regression can be seen as a further development of OLS methodology. For this, a regularization term is added to the sum of squared errors before optimization. The influence of this term is controlled by the hyperparameter λ, which controls the balance between over- and underfitting. The term itself is composed of the sum of the squared estimators. This adds a penalty to the model for the use of estimators with great magnitude. Therefore, ridge regression tends towards shrinking estimators and spreading explanatory power over a broader range of them. As a result, the broader spread of explanatory power makes the model more robust regarding variation in the data. A representation of the optimization problem can be found below. (Rokem & Kai, 2020; Hoque & Aljamaan, 2021; La Tour et al., 2022; Nugroho et al., 2022)

$$\beta = argmin\left(\sum(Y - \hat{Y})^2 + \lambda \sum \hat{\beta}^2\right) \tag{2.6}$$

Advantages of linear regressions are that the relationships they make are easy to understand and to interpret. While a non-linear model will sometimes be able to solve a prediction task in a more accurate way, a linear model will most probably have a more traceable result. Furthermore, linear models are computationally favourable and have advantages for small sample sizes (La Tour et al., 2022).

2.5.2 Gradient Boosting

Gradient boosting represents the idea of combining many weak learners into one strong learner. More specifically, each weak learner is fitted on the residuals of the previous model guided by a loss function. Next, the newly generated weak learner is added to the model, adjusted by a learning rate, creating the next iteration of the strong learner. Therefore, each iteration round is a combination of all weak learners that were fitted in previous rounds. The use of the residuals for each new fit is the reason why the algorithm is called gradient boosting, as the residuals represent the negative gradient of the loss function. After one of the predefined stopping criterions is met, the final gradient boosting model is the sum of all learners adjusted by a learning rate, where the fitting of the learners is guided by a loss

function. As a loss function several options can be chosen and regularization can be added. The weak learners are commonly deployed as decision trees. (Anghel et al., 2018; Fan et al., 2022)

Decision trees answer regression or classification tasks by splitting up the data based on binary criteria until an estimate is reached. The split criteria are determined by splitting the data, iterating through all possible splits, and choosing the best split option based on a loss function. This process is repeated for each created split (node) until a predefined minimal split size or a predefined end point (leaf) sample size is reached. In a classification problem the leaves aim to represent unique classes while in a regression problem the average of the remaining data points is taken. (Xu et al., 2005; Pekel, 2020)

2.5.3 XGBoost

Extreme gradient boosting or XGBoost is a further development of gradient boosting (Fan et al., 2022). For this, solely decision trees are deployed as weak learners and a focus is laid on the scalability of the model. Therefore, one can talk of XGBoost as a gradient boosting model tuned for small process times, application to the biggest possible number of tasks and effective computational resource use. (Chen & Guestrin, 2016; Sahin, 2020)

This is achieved by the implementation of approximate splitting instead of exact greedy splitting. Therefore, not all split possibilities are tested, but percentiles regarding each feature are used to cut computational resource use. Furthermore, shrinking and feature sampling are implemented. These two techniques were not commonly used in other gradient boosting approaches before. Feature sampling conducts a random sampling of available features before split search. Shrinking, on the other hand, introduces another term next to the learning rate that controls the influence of the newly added trees. Additionally, a procedure to handle sparse data was implemented. Conventional tree algorithms are commonly optimized for dense data. This does not reflect the majority of use cases, wherefore the sparse data handling of XGBoost resembles an advantage. Lastly a parallelization, for e.g. split finding across features, and improved memory usage were added to further tune the algorithm. The described improvements are claimed to be the key to the widespread use and success of the algorithm. (Chen & Guestrin, 2016; Sahin, 2020)

2.5.4 MLP

A multilayer perceptron or MLP model is a neural network defined by an input, an output and a varying number of hidden layers. The hidden layers represent the processing between the input and the output. Therefore, they are referred to as “hidden”, as they are not presented to the user of the model. Each hidden layer is composed of a number of nodes. These nodes are fed by all inputs of the layer, adjusted by weights and a node specific bias term. Nodes are also referred to as neurons and the bias is added to influence the activation function independently from the weights. The sum of the weighted inputs, together with the added bias term, is then fed into a non-linear activation function. The result of this function represents the value of the respective node. This process is repeated for all nodes and obtained values become the input of the next hidden layer, including their own weighing, bias and activation. Here, the non-linearity of the activation function becomes imperative, as it enables the model to represent non-linear relationships. Finally, the last hidden layer feeds into the output layer, which presents the result of the model (Itano et al., 2018; Nugroho et al., 2020; Ogunsanya et al., 2023). An example for an MLP model structure can be found in Figure 5.

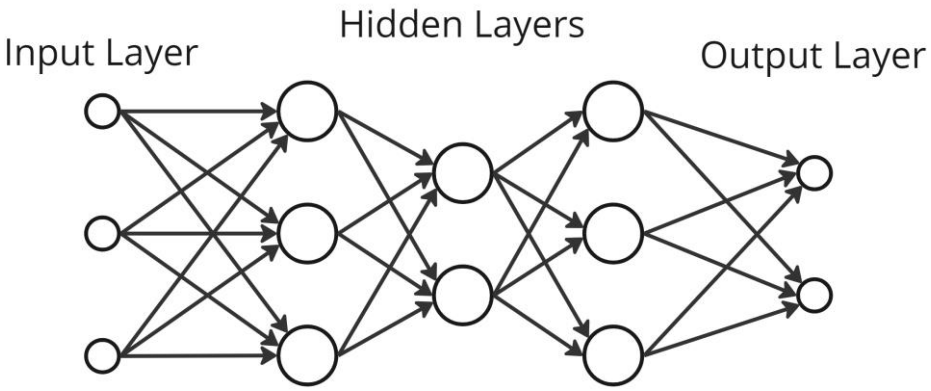

Figure 5: MLP network structure for three hidden layers together with three neurons in the first, two neurons in the second and three neurons in the third hidden layer. (Itano et al., 2018)

Model training is achieved via optimization methods like gradient descent in combination with a loss function to update the weights and biases. Gradient descent indicates the needed direction of change for the weights and biases to minimize the loss function. This process is adjusted by a learning rate to balance over- and underfitting. The described process is called back-propagation. (Itano et al., 2018; Nugroho et al., 2020)

2.5.5 Other relevant models

Other models that are present in this work but were not important parts of it will briefly be introduced on in this section.

The Decision Tree model has a straightforward approach. It makes predictions based on simple split decisions. These split decisions are derived from the features and result in a singular decision tree. Extra Trees, Random Forest, and Bagging Regression are ensemble ML models. These models combine the predictions of multiple decision trees and obtain an outcome by averaging their results. The Bagging Regression performs several bootstrapping rounds and builds a decision tree for each obtained data set. The Random Forest model also performs bootstrapping but introduces variance by only using a randomly selected subset of features to make split decision. The Extra Tree model does not perform bootstrapping but chooses random subsamples from the data. In a next step it builds a decision tree for each subsample with completely random split decisions. (Pedregosa et al., 2011).

K-Nearest Neighbors (KNN) predicts outcomes based on the values of the nearest data points, using their proximity to the input. Elastic Net Regression (ENR) is a linear model that account for multicollinearity and overfitting through regularization. ENR incorporates an additional penalty to perform feature selection by shrinking redundant coefficients to zero (Pedregosa et al., 2011).

2.6 Machine Learning Model Training

A machine learning model must be trained and adapted to carry out classification and regression tasks. Therefore, in the following subchapters machine learning model training and connected concepts will be explained.

2.6.1 Training, Validation and Test Split

To enable the different training phases of a machine learning model, available data must be split into different data sets. The training data set is made to do the actual model training. Model training is composed of adjustments to internal model parameters like weights or data splitting decisions. The validation set is used for hyperparameter tuning and is commonly created from the training data. This is often implemented through cross validation. Lastly, the test set is used to generate an unbiased evaluation of the model performance on data that was not seen during model training or hyperparameter tuning. (Kuhn & Johnson 2013; Yoon, 2021)

2.6.2 Cross validation

Cross validation is a technique that helps to evaluate the performance of trained ML models on unseen data. This is done by repeatedly performing a train validation split for different regions of the data. A common approach is k-fold cross validation. For this, the number of cross validations is defined and the data is divided into the same number of folds. Typical numbers of cross validations are 5 or 10. Afterwards the model training is repeated until each fold was used once as validation data set. Model performance is afterwards reported as an average of the predefined scoring function for all trained models. The advantage of this methodology is that it can accurately identify overfitting and indicates the generalizability of the compiled model. Through the repeated training with different splits, a rather complete image for different particularities is created. Therefore, if specific data points in the training data lead to overfitting, the proneness of the model is revealed through the different training validation splits. Furthermore, the procedure shows how well the model works with different parts of the data acting as unseen data. Therefore, if the model performs equally well in all cases generalizability is shown. (Kuhn & Johnson 2013; Wong, 2015; Berrar, 2019)

2.6.3 Hyperparameter spaces for selected machine learning algorithms

Hyperparameters are settings of machine learning models that are not tuned during training of the algorithms. Therefore, they are higher level tuning options that influence ML model training itself. Accordingly, hyperparameters have to be optimized apart from the model training. (Probst et al., 2019) Hyperparameters and their common tuning ranges for Gradient Boosting, MLP, Ridge Regression, and XGBoost were researched. This was done as these algorithms present important ML models used in this work. Results are presented in Table 3 to 6. Selection of hyperparameters, their workings and the specifics of their tuning are justified in the methodology section.

In Table 3, hyperparameters for Ridge Regression that were found in literature are presented. Sources and areas of application are indicated. Regarding areas of application, an interesting focus of publications on brain activity and health can be observed

Table 3: Hyperparameter ranges for ridge regression. Areas of application are named at first appearance. After that the type of tasks is indicated as classification (C) and regression (R).

Hyperparameter	Hyperparameter Space	Area of application	Source
alpha	(10^{-5}, 10^{15}) - spaced logarithmically, 20 values	Brain activity prediction, regression task	La Tour et al. (2022)
	[0.1, 0.3, 0.5, 0.7, 1, 1.5]	Health prediction of babies after birth, regression task	Nugroho et al. (2022)
	(10^{-4}, $10^{5.5}$, 0.5 $\log_{10}$)	Brain activity prediction, regression task	Rokem & Kay (2020)
	[0.001, 0.01, 0.1, 1, 10, 100, 1000]	Stock price forecasting, regression task	Hoque & Aljamaan (2021)
solver	[svd, cholesky, lsqr, sag, sparse_cg]	R	La Tour et al. (2022)

Hyperparameters, their area of application and respective sources of the MLP algorithm are indicated in Table 4. Focus in the literature is laid on improvement of hyperparameter tuning. Apart from that more literature for classification then regression tasks were found.

Table 4: Hyperparameter ranges for MLP. Areas of application are named at first appearance. After that, the type of tasks is indicated as classification (C) and regression (R).

Hyperparameter	Hyperparameter Space	Area of application	Source
hidden_layer_size	1st layer: [12, 18, 30, 42, 60, 78]	Health prediction of babies after birth, regression task	Nugroho et al. (2022)
	1st layer: (4, 144)	Methodology for improved hyperparameter optimization, classification tasks	Youness et al. (2023)
	1st layer: (10, 15) 2nd layer: (1,10) 3rd layer: (1, 10) 4th layer: (1, 10)	Methodology for improved hyperparameter optimization, classification tasks	El-Hassani et al. (2024)
	1st layer: (1, 16) 2nd layer: (1,16) 3rd layer: (1, 16) 4th layer: (1, 16) 5th layer: (1, 16)	Methodology for improved hyperparameter optimization, classification tasks	Itano et al. (2018)
	1st layer: [3, 6, 9]	Product quality prediction, regression task	Ogunsanya et al. (2023)
alpha	[0.001, 0.005, 0.01, 0.05, 0.1, 0.5]	R	Nugroho et al. (2022)
	(0.0001, 2)	C	El-Hassani et al. (2024)
	(0, 0.001)	C	Itano et al. (2018)
activation	[tanh, relu, logistic]	C	El-Hassani et al. (2024)
	[tanh, rectifier, maxout]	C	Itano et al. (2018)
solver	[sgd, adam, lbfgs]	C	El-Hassani et al. (2024)
learning_rate	[constant, invscaling, adaptive]	C	El-Hassani et al. (2024)
learning_rate_init	(0.001, 0.1)	C	Youness et al. (2023)
	[0.00001, 0.0001, 0.001]	R	Ogunsanya et al. (2023)

Table 5 presents common ranges and value sets of hyperparameters for the gradient boosting algorithm. Reviewed publications focus on improvement of hyperparameter tuning methodologies but also topics regarding health and mining were found. Compared to the previous two algorithms a greater number of hyperparameters was found.

Table 5: Hyperparameter ranges for gradient boosting. Areas of application are named at first appearance. After that, type of tasks is indicated as classification (C) and regression (R).

Hyperparameter	Hyperparameter Space	Area of application	Source
loss	["deviance", "exponential"]	Gold mineral prospectivity mapping, classification task	Fan et al. (2022)
	["deviance", "exponential"]	Image processing for diabetic retinopathy detection, classification task	Datta et al. (2022)
n_estimators	(25, 200, 25)	C	Fan et al. (2022)
	(5, 10)	C	Datta et al. (2022)
	(50, 150)	Methodology for improved hyperparameter optimization, reg. and classification tasks	Young et al. (2018)
	[40, 80, 160, 320, 480]	Methodology for improved hyperparameter optimization, classification tasks	Anghel et al. (2018)
	[200, 400, 600, 800]	Wind energy prediction, regression task	Alonso et al. (2015)
learning_rate	(0.1, 2.1, 0.1)	C	Fan et al. (2022)
	(0.15, 2)	C	Datta et al. (2022)
	(0.00001, 1)	C, R	Young et al. (2018)
	(0.1, 0.3)	C	Anghel et al. (2018)
	[0.05, 0.1, 0.15]	R	Alonso et al. (2015)
max_features	(4, 20, 2)	C	Fan et al. (2022)
	(1, 10)	C, R	Young et al. (2018)
	[0.8, 1] as percentage of all features	C	Anghel et al. (2018)
	[0.3, 0.4, 0.5, 0.6] as percentage of all features	R	Alonso et al. (2015)
subsample	(0.1, 0.8, 0.1)	C	Fan et al. (2022)
	[0.33, 0.5, 0.9]	C	Datta et al. (2022)
	(0.1, 1)	C, R	Young et al. (2018)
max_depth	(2,30,2)	C	Fan et al. (2022)
	[3, 5, 8]	C	Datta et al. (2022)
	(2,10)	C, R	Young et al. (2018)
	[4, 8, 10, 12]	C	Anghel et al. (2018)
	[6, 9, 20, 40]	R	Alonso et al. (2015)
min_impurity _increase	(0,5,1)	C	Fan et al. (2022)
min_samples_split	(0.1, 0.2) as percentage of all samples	C	Datta et al. (2022)
	(2,100)	C, R	Young et al. (2018)
	[2, 4, 8]	R	Alonso et al. (2015)
min_samples_leaf	[0.1, 0.2, 0.3] as percentage of all samples	C	Datta et al. (2022)
	(1,100)	C, R	Young et al. (2018)
	[1, 2, 4]	R	Alonso et al. (2015)

Lastly hyperparameter ranges and value sets for XGBoost are listed in Table 6. A balance between regression and classification tasks can be observed. Apart from that, a wide range of topics is depicted, ranging from landslide prediction, improvement of hyperparameter tuning methodologies to health-related research. As for the gradient boosting, a large number of hyperparameters is provided by the studies.

Table 6: Hyperparameter ranges for XGBoost. Areas of application are named at first appearance. After that, type of tasks is indicated as classification (C) and regression (R).

Hyperparameter	Hyperparameter Space	Area of application	Source
n_estimators	(100, 500)	Prediction of landslide risk, clas. task	Kavzoglu & Teke (2022)
	(100, 600)	Product quality prediction, reg. task	Zou et al. (2022)
	(1, 500)	Methodology for improved hyperparameter optimization, classification and regression tasks	Ørebæk & Geitle (2021)
	(100, 300)	soil property prediction in oil reservoirs based, regression task	Pan et al. (2022)
	[40, 80, 160, 320, 480]	Methodology for improved hyperparameter optimization, classification tasks	Anghel et al. (2018)
learning_rate	(0.1, 1)	C	Kavzoglu & Teke (2022)
	(0.01, 0.3)	R	Zou et al. (2022)
	(0.1, 2)	C, R	Ørebæk & Geitle (2021)
	(0, 1)	Wave run up prediction, reg. task	Tarwidi et al. (2023)
	(0.1, 0.3)	C	Anghel et al. (2018)
	(0.1, 0.3)	Prediction of chronic kidney failure, classification task	Anggoro et al. (2021)
max_depth	(1, 20)	C	Kavzoglu & Teke (2022)
	(1, 10)	R	Zou et al. (2022)
	(1, number of features)	C, R	Ørebæk & Geitle (2021)
	(10, 50)	R	Pan et al. (2022)
	[4, 8, 10, 12]	C	Anghel et al. (2018)
	[1, 3, 5]	C	Anggoro et al. (2021)
min_child_weight	(1, 20)	C	Kavzoglu & Teke (2022)
	(1,9)	R	Pan et al. (2022)
gamma	(0, 0.2)	C	Kavzoglu & Teke (2022)
	(0, 0.05)	R	Zou et al. (2022)
	(0, 5)	R	Pan et al. (2022)
	(0.1, 0.9)	C	Anggoro et al. (2021)
colsample_by _tree	(0.5, 0.7)	C	Kavzoglu & Teke (2022)
	(0.8, 1)	R	Pan et al. (2022)
	(0.8, 1)	C	Anghel et al. (2018)
	(0.1, 0.9)	C	Anggoro et al. (2021)
subsample	(0.8, 1)	C	Kavzoglu & Teke (2022)
	(0.8, 1)	R	Pan et al. (2022)
	(0.1, 0.9)	C	Anggoro et al. (2021)
alpha	(0.01, 0.3)	R	Zou et al. (2022)
	(0, 0.2)	R	Pan et al. (2022)
lambda	(0, 1)	R	Zou et al. (2022)
	(0, 0.2)	R	Pan et al. (2022)

2.6.4 Grid search

Grid search is a methodology to guide the application of hyperparameter tuning. During grid search , a predefined set of values for each hyperparameter is used and a grid of all possible combinations is compiled. Afterwards, each of the combinations is tried out and the best performing set of hyperparameters is selected. To achieve this selection, an upfront determined scoring function is used. The advantage of this methodology is that each combination is tried out, wherefore no optimum present in the grid can be missed. On the other hand, the method is computationally expensive and not flexible. This means that an optimum could be missed, if it lies between two grid points. Furthermore, through the try-out of each possible combination, many irrelevant options will be tested. Nevertheless, grid search resembles the most commonly deployed hyperparameter tuning methodology. (Bergstra & Bengio, 2012; Youness et al., 2023)

3 Knowledge gap and research question

In the following chapter, the knowledge gap for this thesis is derived. This is done based on the findings from the introduction and the state-of-the-art section. With the knowledge gap as foundation, the research question will be formulated to aid the generation of the sought-after insights.

3.1 Knowledge gap

In the previous chapters, it was shown that improved material separation can help to reach EU recycling goals and to diminish CO_2 eq. emissions. Regarding CO_2 emission savings, promising annual savings of up to 200 million tons of CO_2 eq. for Europe were identified in the introduction. Furthermore, indications for economic advantages for waste separation plant proprietors were found. To understand what is required to achieve these economic and environmental benefits, shortcomings of nowadays waste separation plants were analysed.

The optimization of machine settings and machine interplay in waste separation plants was identified as main hindrance. Waste separation plants are commonly only optimized once during their commissioning. This is due to the cost of probing and little knowledge regarding waste stream composition. Furthermore, state-of-the-art waste separation plant modelling mainly relies on expert valuation and experience values for separation efficiencies. This is problematic due to two reasons. First, plant modelling and process characterization become static. This means that separation efficiencies cannot be adopted, as information about inputs and composition of waste streams in the plant is missing. Second, waste composition is constantly changing, sometimes even per season. Therefore, plant optimization is outdated rather quickly, leading to suboptimal separation results. To avoid this, more frequent optimization is needed.

Delving into requirements for improved plant optimization, another hindrance was identified: separation efficiencies cannot be transferred easily from one plant to another. This is due to the fact that waste composition changes based on regionality, wherefore generalizability is hampered. Therefore, plant specific process parameters like separation efficiencies and area densities are required to enable real-time plant optimization. To achieve this, data about material stream composition in the waste sorting plants is needed. Nevertheless, a sufficient coverage of sensor-based waste stream classifiers is prohibited from an economical perspective.

To still obtain a reasonable process characterization, the use of data from sensor-based sorters that are already present in the plant is proposed. Examples of this can be found in works of Küppers et al. (2020) and Kroell et al. (2021, 2024a, 2024b). In their studies, they determined separation efficiencies of NIR-scanners based on material occurrence and occupancies and achieved reasonable accuracies. Furthermore, weight-based quality control, with the help of a NIR-scanner and laser triangulation, could be showcased.

As a shortcoming, the described studies were conducted with fully characterized material. While this is favourable for the explanatory power of the results, implementing these approaches into a real-world waste sorting plant would require extensive probing. This probing would need to be conducted on a regular basis and is therefore prohibited from an economic viewpoint. Another disadvantage is that for some studies machinery was modified in way that is normally not present in commercial waste sorting plants. An example for this would be the addition of laser triangulation for height detection of particles to a NIR-scanner. Lastly, the discussed studies showcase dynamic separation efficiency determination and waste stream characterization as stand-alone concepts but do not research their interplay.

This study, therefore, tries to implement material characterization in an industrial setting with data supplied from a waste sorting plant in Scandinavia. This means that no fully characterized material is available for algorithm training and testing, which resembles real world conditions. Furthermore, a dynamic material separation efficiency determination is implemented. This will be done in the form of material stream prediction from one NIR-scanner to the other. These methodologies will be applied together to predict the purity of the HQ agglomeration product. Therefore, not only concepts introduced by Kroell et al. and Küppers et al. will be applied to a real-world sorting plant, but also their interplay is researched.

Further novelty of the research is given through the use of belt weighers data. No studies regarding belt weigher data use for waste stream characterization in plastic sorting plants could be found during literature research. Therefore, insights in the opportunities arising from the use of this data will be obtained. Apart from that, this study contributes to the literature by using NIR-scanner data for material stream characterization, which is originally produced for sorting of the materials. The benefits of the use of this data are described and showcased by several authors, but through the application in an industrial setting, new insight can be generated.

3.2 Research question

Concluding from the introduction, the state-of-the-art section and the identified knowledge gap the following research question can be formulated:

How can the quality of the agglomeration line product in a plastic waste separation plant be predicted based on NIR-scanner and belt weigher information through a data driven approach?

This research question tries to facilitate the aspirations of the goal of this study and the derived knowledge gap. Through the quality prediction of the agglomeration line product the interplay of dynamic material separation and area density determination can be showcased. As a smallest possible building block, the answering of this research questions can showcase concepts and methodologies that enable real-time process optimization for waste sorting plants. To answer the main research question, the following sub research questions are compiled:

- *What correlations and relationships exist in the data?*
- *How can the area density, the area flow prediction and the separation efficiency be modelled?*
- *How does a joint application of the developed concepts perform for quality prediction of the agglomeration line product?*

The first sub-research question guides the exploration of the data that is available in the waste separation plant. This is necessary to identify modelling opportunities and to assess the quality of the data. Understanding of this is important to generate insights about the uncertainty that comes with the application of the prediction and modelling methodologies. Furthermore, the gained knowledge will be applied to guide data pre-processing. The second sub-research question provides the setup for the modelling techniques that are tried out with the explored data. Here, the required information, that was identified through the knowledge gap, is gathered. Furthermore, the performance of the model building blocks is assessed. The third sub-research question guides the analysis of the interplay between area density and separation efficiency modelling as well as the connections to the area flow prediction. This is important to move from stand-alone observations to a joint analysis of opportunities for waste stream characterization and prediction.

4 Methodology

To answer the research question, a methodology to tackle the established knowledge gap was set up. The governing approach consists out of a data exploration phase, statistical modelling for area density determination, ML modelling for separation efficiency determination, area flow prediction and a joint application of all three concepts. A visualization of the different phases can be found in Figure 6.

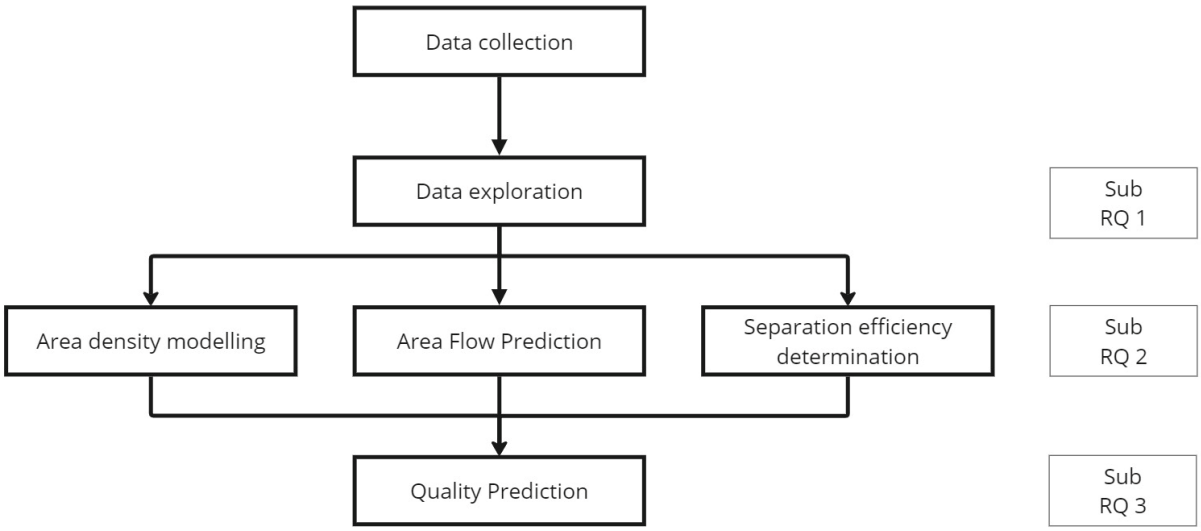

Figure 6: Overarching methodology for the answering of the research question, different phases and connected sub research questions are indicated.

The data exploration is connected to sub research question one, to search for already existing correlations and relationships in the provided data. Area density modelling, area flow prediction and separation efficiency determination are addressing sub question two. This connection is made, as they represent the modelling and calculation of important information necessary to answer the overall research question. Sub research question three is represented through the final quality prediction. During quality prediction, all modelled concepts are applied together, wherefore their interplay can be probed and analysed.

4.1 Used software

To conduct calculations, modelling and data handling Python 3.11.7 in the integrated development environment Spyder 5.5.1 was used.

Data handling was performed with the help of the Pandas 2.14 library. Basic mathematics and indicator calculation were implemented though the use of Numpy 1.26.4. Apart from that, the Statsmodels 0.14.0 library aided the implementation of the OLS modelling, while the Scikit-learn library 1.2.2 was used for machine learning applications. Finally, plotting was conducted with the help of Matplotlib 3.8.0.

4.2 Indicators and Metrics

Several indicators and metrics were used during the thesis and will be explained in this subchapter.

The Pearson correlation coefficient is a measure for the linear correlation between two data sets. In the remainder of this work it is denoted as r. For a perfect positive correlation, the coefficient will be 1 while for a perfect negative coefficient it will result in -1. If no correlation at all is present, the Pearson correlation coefficient will be calculated with 0 (Cohen et al., 2009; Rainio et al., 2024). A mathematical formulation of the Pearson correlation coefficient can be found below, with x_i representing a value of the first data set and $\bar{x}$ representing the mean of the first data set. The same notation is applied to the second data set with y as its representation.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4.1}$$

The mean absolute error (MAE) is a useful metric to determine the mean deviation from predicted to measured values. It is determined by summing the absolute errors of the model and then dividing it by the number of samples. Due to this calculation, it does not lay specific emphasis on outliers or other special cases in data. Therefore, it can be considered rather robust (Pedregosa et al., 2011; Rainio et al., 2024). The formula of the MAE can be found below, where y_i represents one of the observed values, $\hat{y}_i$ indicates the prediction of the same observation, Y is the vector of all observed values, $\hat{Y}$ represents the vector of all predicted values and $n_{Samples}$ is the number of samples.

$$MAE(Y, \hat{Y}) = \frac{1}{n_{Samples}} \sum_{i=1}^{n_{Samples}} |y_i - \hat{y}_i| \tag{4.2}$$

The mean squared error (MSE) is determined by squaring all errors of the model, summing them up and dividing the result by the number of samples. Therewith, the mean squared deviation from the predicted to the measured values gets computed. Due to the squaring, attention is given to errors of bigger magnitude, as they gain greater influence on the result (Pedregosa et al., 2011; Naidu et al., 2023; Rainio et al., 2024). A mathematical formulation of the MSE is indicated below, where y_i and $\hat{y}_i$ are single values from the observation vector Y and the prediction vector $\hat{Y}$. Lastly $n_{Samples}$ indicates the number of samples that were taken.

$$MSE(Y, \hat{Y}) = \frac{1}{n_{Samples}} \sum_{i=1}^{n_{Samples}} (y_i - \hat{y}_i)^2 \tag{4.3}$$

The coefficient of determination, also denoted as R^2, measures the goodness of fit of a model. It indicates how well the model's predictions resemble the observed values. Putting it into other words, R^2 represents the proportion of variance, in the measured data, that is explained by the model's predictions. R^2 can have values between 1, indicating the perfect fit, and minus infinity, as the model can be arbitrarily worse. A model that always predicts the mean of the data results in 0 and can, therefore, be seen as baseline. To calculate R^2, the sum of the squared difference between measurement and prediction is divided by the sum of the squared difference between each measurement and the mean of the measurements. The result of this term is then subtracted from 1 and R^2 is obtained (Hagquist & Stenbeck, 1998; Pedregosa et al., 2011). On the next page, a mathematical representation of R^2 is depicted, where Y is the vector of all observed values, $\hat{Y}$ is the vector of all predicted values and y_i and $\hat{y}_i$ represent single values from these vectors.

$$R^2(Y,\hat{Y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4.4)$$

To predict the quality of the agglomeration product, a weight percentage metric is necessary. For this, the material of interest is divided by the total material. This indicates the weight share of the material in the product. A mathematical formulation of the indicator can be found below, where $m_{material}$ represents the weight of a specific material and m_{Total} indicates the weight of all materials that are present.

$$wt\%_{Material} = \frac{m_{Material}}{m_{Total}} \quad (4.5)$$

4.3 Data Composition, Collection and Processing

In this subchapter, the retrieval of the data that was used in this work will be explained. Furthermore, emphasis is laid on the composition of the data and initial processing that enabled data exploration.

NIR-scanner data was retrieved through the TOMRA Insight portal and by an influx database provided by Bremen University. The influx database aggregates data from the TOMRA Insight portal and facilitates downloads of greater time intervals. Here, the finest available resolution is one data point per minute. Data was retrieved in the unified material statistic format. This format was developed to ensure comparability between different TOMRA NIR-scanner units. Statistics that were retrieved show material-specific area flows as well as the placement of the material on the conveyor belt in [m^2/h]. Available material categories are shown in Table 7. Although, the unified material statistic is made for comparability and to ensure similar material categories across all scanners, BOARD_CT and PET_G are missing for PO75C. Additional data like the sorting program, valve cycles and lamp health are available but have not been used for the conducted research. Nevertheless, this data could be used in the future.

Table 7: Material categories for PO75 and PO75C indicated with their original name, the name used in this work, their meaning and the sorting indication per scanner.

Unit	Original Name	Meaning	Name in this work	Eject	Drop
PO75	BOARD_CT	Corrugated Cardboard			X
	EPS	Expanded Polystyrene			X
	GT	Getränke Karton (= Beverage Carton)	BC		X
	OTHER_POLYMERS	All polymers that do not have an own category			X
	PAPER	Paper			X
	PE_FILM	Films of Polyethylene		X	
	PE_RIGID	Objects of Polyethylene that are not film		X	
	PET_BOTTLE	Bottles made of Polyethylene terephthalate			X
	PET_G	Glycol-modified Polyethylene terephthalate			X
	PET_MONO_TRAY	Polyethylene terephthalate only trays			X
	PP	Polypropylene		X	
	PP_FILM	Films of Polypropylene		X	
	PS	Polystyrene			X
	PVC	Polyvinyl chloride			X
PO75C	EPS	Expanded Polystyrene			X
	GT	Getränke Karton (= Beverage Carton)	BC		X
	OTHER_POLYMERS	All polymers that do not have an own category			X
	PAPER	Paper		X	
	PE_FILM	Films of Polyethylene			X
	PE_RIGID	Objects of Polyethylene that are not film			X
	PET_BOTTLE	Bottles made of Polyethylene terephthalate			X
	PET_MONO_TRAY	Polyethylene terephthalate only trays			X
	PP	Polypropylene			X
	PP_FILM	Films of Polypropylene			X
	PS	Polystyrene			X
	PVC	Polyvinyl chloride		X	

After retrieval, the data was cleaned from missing values and brought into a multilevel column index form. Afterwards, the data was sorted by unit, statistic and belt part as well as material or drop or eject stream if applicable.

Regarding the belt weighers, data was retrieved from Sutco through a MySQL database, with the help of a SECOMA connection. Data is available in 3-7 second steps and was aggregate to the same time steps as the NIR-scanner data.

In total, four-month worth of data were used. The data set was recorded from the second of March 2024 at 12:35 to the second of July 2024 at 7:11. To avoid a temporal offset between the different data sources, the data was synchronized. This was done with the help of the Pearson correlation coefficient. Temporal offsets can e.g. arise due to unaligned time stamps within the machines or through the time that the material needs to travel from one unit to the next.

4.4 Data Exploration

To assess the model building capabilities of the data, exploration of the data is necessary. Avoidable noise and special events that could hamper explanatory power have to be identified. Proper exploration enables to either remove disruptive data points or to develop strategies to treat these cases.

The data exploration for belt weigher data, NIR-scanner area flows and belt occupation data was guided by the Pearson correlation coefficient and the analysis of the data distribution. The latter was done from a frequency view point. Here, high correlations and normal distributions were sought of, wherefore data pre-processing strategies were developed based on this regard.

4.5 Statistical modelling – Area densities

To determine the area densities, ordinary least squares (OLS) modelling was applied. This was done to enable the conversion from area flow data to mass flow data and ultimately facilitate quality prediction in weight percent. Since the connection between area flow and mass flow through area density is linear, OLS modelling was chosen. This was done because OLS is a linear modelling method, which fits the introduced relationship. Apart from that, an 80/20 training test split for model training and testing was applied.

After application of the model, outcomes from Statsmodels were checked for violations of assumptions. Additionally, signs of invalid model properties were examined and discussed.

4.5.1 Multicollinearity

Multicollinearity was tested and detected according to Shrestha (2020). This was done with the help of a correlation matrix. In a next step, multicollinearity was treated through grouping of independent variables as proposed by Paul (2006).

4.5.2 Bootstrapping

Bootstrapping was applied to gain additional trust in the obtained area densities. This was done next to the confidence intervals obtained from the initial modelling. For application, 10,000 bootstrapping rounds for each area density were conducted. From this, area density distributions were determined, along with quantiles similar to the confidence intervals from the initial modelling. Additionally, the mean and median of the distributions were calculated.

4.5.3 Testing and evaluation

To test the obtained area densities, they were multiplied with the area flows from the test set. Afterwards they were summed up and plotted against the belt weigher data of the same time frame. For evaluation the MAE was used. Additionally, generalizability of the area densities to other scanners was tested.

4.6 Machine learning – Area flow prediction

For area flow prediction, ML models were identified as appropriate approach. This judgement is based on experience from the Sutco research and development department. They found out that for the prediction of area flows from one scanner to the next scanner no purely linear relationships are present. Therefore, non-linear modelling is needed. Furthermore, physical properties of waste are changing and limited information on waste characteristics is available through the limitation on 2D data. Therefore,

machine learning was preferred over a physics-based approach or empirical models. For the ML modelling, an 80/20 training test data split was applied.

4.6.1 Model Try-Out and selection

As proposed by Kuhn & Johnson (2013), a broad range of models was tried out. Furthermore, a variety of ML model architectures was represented. Accordingly, the following models were tested: Decision Tree, Extra Tree, Random Forest, Gradient Boosting, K-nearest Neighbours, Bagging Regressor, Ridge Regression, Elastic Net Regression and MLP. Regarding model architectures the Decision Tree model represents a simple decision tree architecture. Extra Tree, Random Forest, Bagging Regressor and Gradient Boosting resemble tree-based ensemble models, where a combination of various trees leads to the prediction outcome. K-nearest Neighbours is a neighbour-based regression algorithm where the predicted value is computed as a function of the measured values of the input's nearest neighbours. Ridge Regression and Elastic Net Regression represent linear regression models. Lastly, MLP is a neural network type machine learning model. (Pedregosa et al., 2011)

Models are selected according to their performance and their simplicity, as presented by Kuhn & Johnson (2013). Balancing these two properties keeps computational expenses at bay and improves interpretability. The interpretability of simpler models is given, as their final model structure tends to stay close to real world principles, instead of being based on decision trees or nested mathematical functions.

4.6.2 Hyperparameter tuning

Hyperparameter tuning was conducted based on grid search and applied to the training data set. This was done to leave the test data set for final performance evaluation.

To scope down the selection of hyperparameters to a reasonable computational effort, a maximum of five hyperparameters per ML model was selected. Selection was guided by the amount of mentions. Furthermore, areas of application that included regression task were prioritized, as they better reflect the task at hand. Selected hyperparameters and their ranges are described in the following, together with a small explanation of their workings.

For ridge regression, two hyperparameters were found in literature and applied accordingly. These hyperparameters are the solver of the regression and the alpha value.

The solver hyperparameter is responsible for the computational implementation of the underlying mathematics. Nevertheless, the hyperparameter was considered to indicate which of the solvers is the fastest. For hyperparameter selection all options were taken over, giving the following range for the grid search: [svd, cholesky, lsqr, sag, sparse_cg]. (Pedregosa et al., 2011)

The following values were set for the alpha hyperparameter: [0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 500, 1000, 5000, 10000]. This range was chosen to include the default parameter value and to represent the widespread range of values found in the literature. The alpha hyperparameter controls the influence of the sum of squared weights added to the loss function in the ridge regression model, affecting the penalty applied to high magnitude weights. (Pedregosa et al., 2011; La Tour et al., 2022)

In Table 8, a summary of the hyperparameters and their respective values of the Ridge Regression model can be found.

Table 8: Hyperparameter sets for ridge regression applied during grid search.

Hyperparameter	Values applied in grid search
alpha	[0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 500, 1000, 5000, 10000]
solver	[svd, cholesky, lsqr, sag, sparse_cg]

For the MLP model, hidden_layer_size, alpha, activation and learning_rate_init were selected as hyperparameters. It was decided to try out 9 hidden_layer_size versions, as it resembles the most important hyperparameter of the algorithm. Therefore, the other hyperparameters have been limited to three options each. This was done to keep the grid search computationally reasonable.

The hidden_layer_size hyperparameter determines the number of hidden layers and the respective number of nodes. An increased number of nodes enables the layer to represent more complicated relationships between input and output. Looking at the number of layers, an increased number of hidden layers enables the detection of "deeper" patterns. This is achieved through the repeated use of the activation function and updated weighing. Therefore, each layer refines the pattern from input to output (Pedregosa et al., 2011; Itano et al., 2018). For the hyperparameter, a use of one to three hidden layers was chosen. To not overcomplicate things, all layers have the same size with either 50, 100 or 150 nodes per layer. The number of nodes is slightly higher than seen in literature, but was chosen to balance out the decreased number of layers.

The MLP model adds the sum of squared weights to the result of its loss function. The influence of this term is controlled by the alpha hyperparameter (Pedregosa et al., 2011; La Tour et al., 2022). For hyperparameter tuning, the set of alpha values was chosen with: [0.001, 0.01, 0.1]. This was done to represent the values found in literature. Interestingly, all encountered values were higher than the default value of 0.0001 but kept below 1 for the majority of the findings.

The activation hyperparameter selects the activation function that is applied to each node of the hidden layers (Pedregosa et al., 2011; Nugroho et al., 2020). For this, "tanh", "relu" and "logistic" were selected, as they are the three applicable activation functions to the problem at hand. Apart from that the "identity" function is available. This function is commonly used when near-linear behaviour is expected, as it does not modify the input. (Pedregosa et al., 2011)

An overview of hyperparameters and their ranges that were applied in the MLP tuning process are depicted in Table 9.

Table 9: Hyperparameter sets for MLP applied in grid search.

Hyperparameter	Values applied in grid search
hidden_layer_size	1 layer: [50, 100, 150] 2 layers: [[50, 50], [100, 100], [150, 150]] 3 layers: [[50, 50, 50], [100, 100, 100], [150, 150, 150]]
alpha	[0.001, 0.01, 0.1]
activation	[tanh, relu, logistic]

As hyperparameters for the gradient boosting model, n_estimators, learning_rate, max_features, max_depth and min_sample_split were selected. Min_sample_split and min_sample_leave have equal amount of mentions, but due to their similarity max_sample_split was chosen. This was done, as the decision to continue splitting or not, was preferred to be done before the split rather than after the split. Apart from that, the number of options per hyperparameter were limited to three. With this setting, a total of 243 combinations was tested during grid search, which represents a reasonable use of computational power.

The n_estimators hyperparameter dictates how many boosting stages are performed. This means that it controls how often a new tree will be fitted to the residuals of the latest model (Alonso et al., 2015; Fan et al., 2022). For n_estimators, grid search values are set to [50, 250, 500]. This was done to test for values that are below and above the default value of 100 and to represent the range of values found in literature.

The impact of each added tree on the overall model performance is determined by the learning rate (Pedregosa et al., 2011; Alonso et al., 2015; Fan et al., 2022). Learning rate values are set to [0.05, 0.15, 0.3]. This was done to explore options around the default value of 0.1.

To adjust the maximum number of features that are used for split calculation the max_feature hypermparameter is tuned (Pedregosa et al., 2011; Datta et al., 2022). Following the percentual approach found in literature, together with a number of 21 features, the max_feature values were set to [7, 14, 21].

The max_depth hyperparameter indicates the maximum level of layers in a decision tree. This directly influences its complexity (Pedregosa et al., 2011; Datta et al., 2022). With a default of 3, the values for max_depth were set to [2, 10, 18]. This was done to check options below and above the default values and to represent the greater values found in literature.

The decision of how many observations must be contained in a node to make a split is made by the min_samples_split hyperparameter (Pedregosa et al., 2011; Alonso et al., 2015). The values for the hyperparameter were set to [2, 50, 100]. This was done to have the default value of 2 within the range and to test higher values present in the literature.

In Table 10 all hyperparameters and their ranges that were used for Gradient Boosting hyperparameter tuning are depicted.

Table 10: Hyperparameter sets for gradient boosting applied in grid search.

Hyperparameter	Values applied in grid search
n_estimators	[50, 250, 500]
learning_rate	[0.05, 0.15, 0.3]
max_feature	[7, 14, 21]
max_depth	[2, 10, 18]
min_samples_split	[2, 50, 100]

Hyperparameters that were chosen for XGBoost hyperparameter tuning are n_estimators, learning_rate, max_depth, gamma and lambda. Alpha and lambda have the same amount of mentions and both resemble regularization terms for the loss function. As lambda represents a preferred regularization method, it was chosen for grid search instead of alpha. Furthermore, the number of

values per parameter was limited to three. This was done to not increase the number of possible combinations above the limitation that was set for the other models.

The n_estimators hyperparameter selects the total number of times a new tree gets fitted to the residual of the model. High number of rounds can lead to overfitting and unnecessary expense of computing power. On the other hand, a low number of rounds could have underfitting as a consequence (Pedregosa et al., 2011; Kavzoglu & Teke, 2022). The following values were used for the grid search: [50, 175, 300]. This was done to represent the range of values found in literature, while avoiding emphasis on the extremes.

To influence the learning rate of XGBoost, the learning_rate hyperparameter is used. This parameter determines the influence of each newly added tree to the model (Tarwidi et al, 2023). Here, extensive learning rates can oversee optimal solutions, while too conservative learning rates can trap the model into a local optimum (Itano et al., 2018). For the learning rate, the subsequent set of values was selected: [0.1, 1, 2]. This was done to resemble the range of values used in literature with a focus on regression tasks.

The max_depth hyperparameter sets the maximum number of splits a tree branch can have until reaching an end point. Higher values allow the model to represent more complicated relationships but come with the risk of overfitting (Pedregosa et al., 2011; Kavzoglu & Teke, 2022). The following set of values was selected for the max_depth parameter: [1, 8, 15]. With this, the minimum value from the literature and greater values that were found for regression tasks are tested.

Gamma represents the minimum loss function reduction that is needed to further split the data (Chen & Guestrin 2016; Tarwidi et al, 2023). The hyperparameter prevents overfitting and can be used to safe computational resources by stopping model training when no significant improvement is detected. Seeking orientation in literature with a focus on regression tasks, the following values were chosen: [0, 0.25, 0.5].

The lambda hyperparameter has the same workings as the alpha hyperparameter for the MLP and the Ridge Regression model. Finding a middle ground between the values from the literature, the following values were selected: [0, 0.15, 0.3].

A summary of the hyperparameters that were used for the tuning of the XGBoost model can be observed in Table 11.

Table 11: Hyperparameter sets for XGBoost applied in grid search.

Hyperparameter	Values applied in grid search
n_estimators	[50, 300, 550]
learning_rate	[0.1, 1, 2]
max_depth	[1, 15, 30]
gamma	[0, 0.25, 0.5]
lambda	[0, 0.15, 0.3]

4.6.3 Cross validation and grid search

A 5-fold cross-validation was applied during grid search to test the generalizability of the ML models. Furthermore, this approach was chosen to increase robustness against outliers during model selection. Model selection was conducted based on the mean of the indicators obtained from the cross validation. As guiding scoring function the MAE was used.

4.6.4 Testing and evaluation

To test and evaluate the area flow prediction, the test set was used and predictions were plotted against measured data. Evaluation was guided by the MAE, MSE and R2 only played a secondary role.

4.7 Separation efficiency

Separation efficiencies from PO75C to the agglomeration product were taken over from the sorting step between PO75 and PO75C. This was done as no validation data for the last separation step was available. Validation data is needed for model training and testing. Therefore, the assumption of transferable separation efficiency was made to showcase the full solving of the research question. When a suitable modelling technique without validation data needs is identified, or validation data is obtained, this part of the work should be revisited.

The absence of validation data is given, as no final quality determination is carried out in the process. Insufficient quality is normally reported by the buyer of the product to the operator of the plant. This comes with the difficulty of mapping the reported bale to a temporal scope of processing in the plant. Furthermore, there is only a very limited amount of data available, if data is recorded at all.

To determine the separation efficiency, it was assumed that separation efficiencies are constant over time and that ejected material separates equally well. Additional focus was laid on material that is ejected together with the targeted material. The accidently ejected material was ascertained as percentual share of the separated target material. Apart from that, separation efficiencies were determined based on the area flows instead of mass flows. This was done to not include additional uncertainty from area density determination into the calculations.

4.8 Quality prediction

To conduct the final quality prediction, the area density modelling, the area flow prediction and the separation efficiency determination were applied together. Here, the area flows on PO75C were predicted by the area flows on PO75. Next, the separation efficiencies were used to determine the area flows after the last separation step. Lastly, the predicted area flows for the agglomeration product were converted into mass flows using the modelled area densities. These steps enable quality prediction in weight percent which is required for quality determination.

Evaluation was conducted between results that were obtained with predicted area flows and results that were realized with the help of measured area flows. As main indicator the MAE was used. The obtained mass flows include several sources of uncertainty. For PO75, the uncertainty comes from area density modelling. For PO75C, the material-specific mass flows include the uncertainties from the area density modelling and the area flow prediction. Finally, for the agglomeration product, the uncertainty arises from area density modelling, area flow prediction, and separation efficiency determination.

5 Results and discussion

In the following chapter, the results of the conducted research will be presented and discussed. As there will be frequent reference to specific units in the plant, in Figure 7 a repeated representation of the observed plant part can be found.

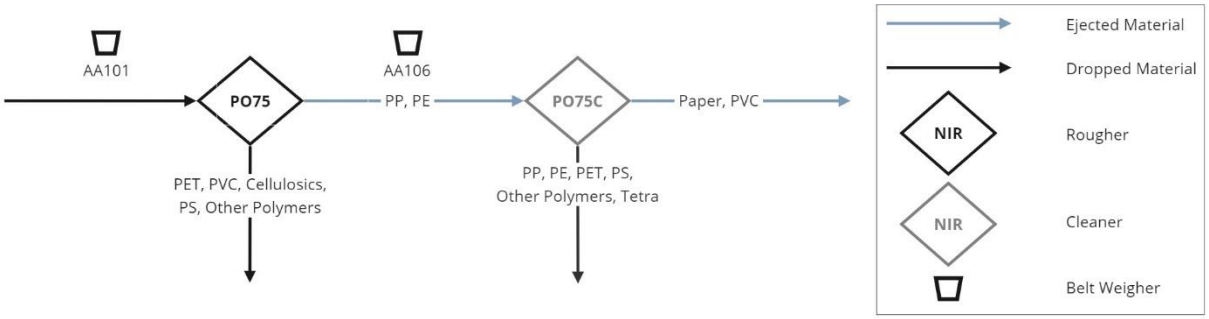

Figure 7: Schematic of the observed part of the agglomeration process in the waste separation plant.

The belt weigher that measures the input of the first scanner is referred to as AA101, while the belt weigher measuring the eject stream is indicated as AA106. The first separation step is conducted by the PO75 rougher and the second separation step is achieved through the PO75C cleaner. Upfront and following there are many more separation and processing steps, but it was focussed on this part due to the high availability of data. Furthermore, this was done to narrow down the number of separation units to a reasonable amount for the scope of this thesis. What is also noticeable that for PO75 the eject stream is of interest, as this is the stream that gets fed into the cleaner. On the other hand, for PO75C the dropped stream is of interest, as it resembles the product.

5.1 Data exploration

Data exploration was done to delve into the characteristics of the data and to ease application of methodologies important to this research. For this, special focus was laid on sanity checks, correlation between data sets and the distribution of the data. This information helps to generate insights about uncertainty in the data. Understanding this uncertainty is important, as it will influence the explanatory power of the established quality prediction. Through this subchapter, sub research question one will be answered, as correlations and relationships in the available data are researched.

5.1.1 Belt Weigher Data

Starting off, the distributions of the belt weighers AA101 and AA106 were plotted and can be found in Figure 8. On the x-axis the measured weight in [t/h] is displayed, while the y-axis indicates the number of occurrences over the observed time frame.

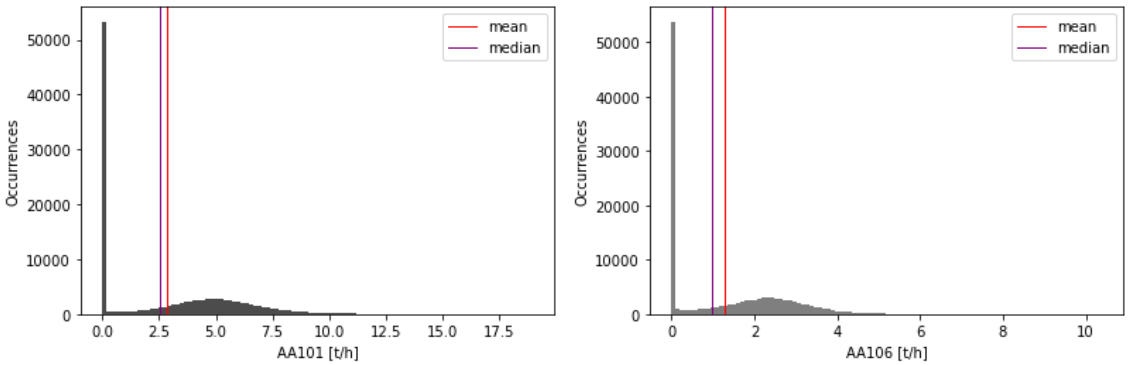

Figure 8: Mass flow distribution in [t/h] for belt weighers AA101 and AA106 without further processing.

Mass flow ranges from 0 [t/h] to 18.97 [t/h] for AA101 and from 0 [t/h] to 10.37 [t/h] for AA106. Due to the heavy left tail in the data, it can be claimed, that the majority of the data is empty. Therefore, zeros have been removed. Nevertheless, after the removal no significant change was observed. This is due to the fact that the belt weighers tend to measure very small weight flows, although the belt is empty. To balance this out, several thresholds for removing the left tail were tested. As a result, excluding data below 0.3 [t/h] was considered an appropriate threshold. Outcomes of this procedure can be examined in Figure 9.

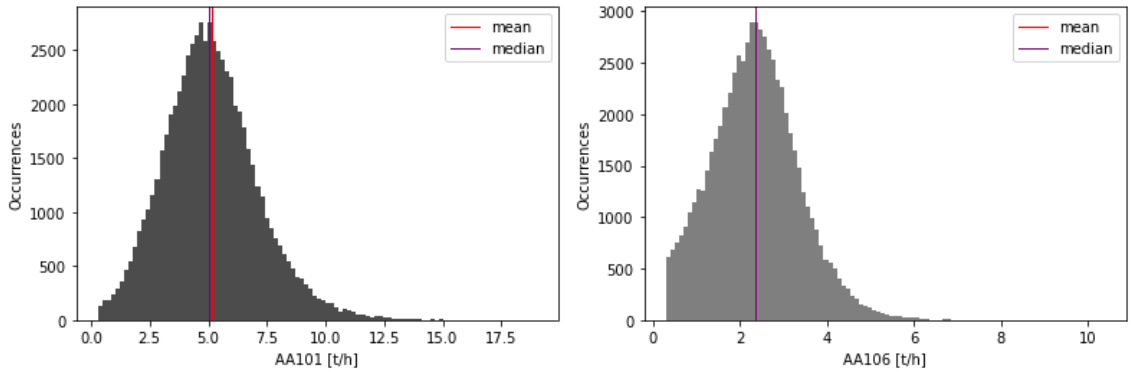

Figure 9: Mass flow distributions in [t/h] for AA101 and AA106 after removal of all values below 0.3 [t/h].

Nearly normal distributed mass flow distributions can be observed in the newly generated plots. The mean for AA101 is determined with 5.14 [t/h] and with 2.37 [t/h] for AA106. Medians are rather similar with values of 5 [t/h] and 2.34 [t/h]. Belt weigher AA106 measures the mass flow after the first scanner (PO75). This also resembles the input to PO75C. As between PO75 and PO75C a separation step happens, the more than halved mean from AA101 to AA106 is a good sign of consistency in the data. Nevertheless, the distribution is cut on the left side for AA106. This is due to the fact that, for AA101, the threshold had to be moved up to 0.3 [t/h] to eliminate the vast amount of empty data. Naturally, the same data was removed for AA106, as the data sets are used in combination. A possible explanation for this difference is that AA106 has lower expected mass flows. Therefore, the machine could be set to a greater sensitivity. On the other hand, for AA101, greater mass flows are expected, wherefore the sensitivity could be toned down to handle higher weights more accurately.

In a next step, the correlation between AA101 and AA106 was tested. For this, a scatter plot before and after the removal of the low value tail was compiled and can be found in Figure 10. On the y-axis, mass flows measured by AA106 are presented and on the x-axis mass flows recorded by AA101 are indicated.

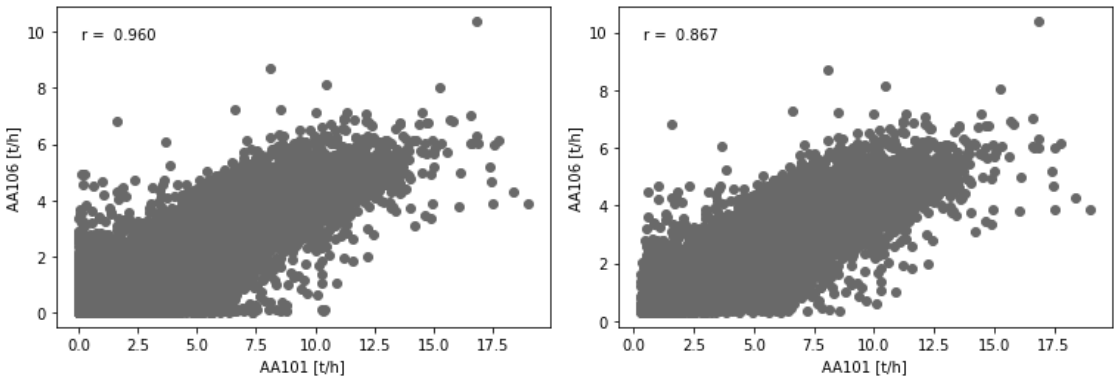

Figure 10: Scatter plots between belt weigher AA106 and belt weigher AA101, before the cut of low-value data (left) and after the removal of low-value data (right). The pearson correlation coefficient is indicated at the top left corner of both plots.

What can be observed is that r decreases after the removal of the low-value tails. At the same time, the amount of impossible values decreases. These values are present on the left-hand plot, when one of the axes has values of zero, but the other belt weigher is still detecting a mass flow. This is most probably due to a temporal measurement error of one belt weigher, while the other belt weigher is functioning. The decreasing r can be explained by the removal of the close to zero values. In the left-hand version of the plot a vast amount of data is empty for both belt weighers. When both belt weighers measure values close to zero, the correlation of the respective value pairs is high. This effect could explain the higher correlation before the removal of the low value tail.

To gain an advanced understanding of the correlation of the belt weighers, the data was split in 200 bins by time. In a next step, r was computed for all bins. Regarding plots can be found in Figure 11.

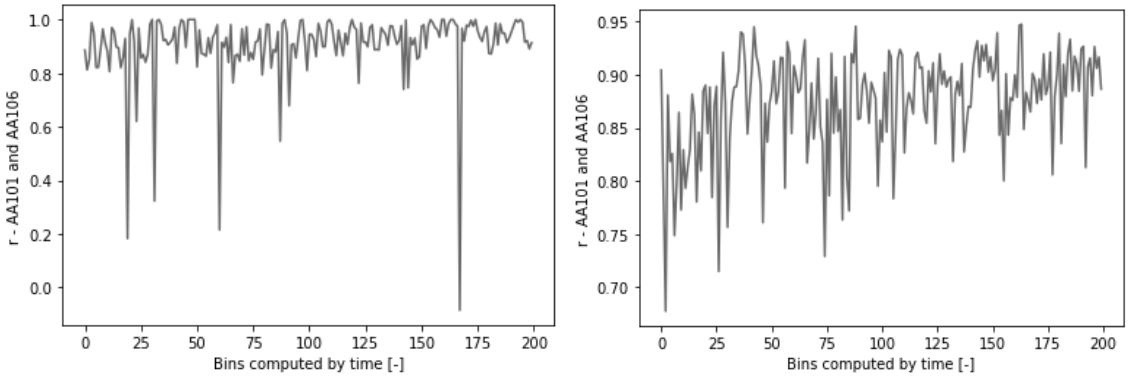

Figure 11: Pearson correlation coefficient for belt weigher AA101 with regards to belt weigher AA106 with 200 bins compiled by time, before the removal of low-value data (left) and after removal of low-value data (right).

Before the removal of low value data, correlation can go down until a r value of -0.09 but also reaches a value of 1 on a frequent basis. After the low-value data is eliminated, the extremes are removed from the correlation plot. Maxima of the right-hand plot are reached with values of 0.95, while minima go

down to values of 0.68. Positive extremes of the first plot can again be explained with the values close to zero. Although they represent faulty datapoints, they are still very similar in magnitude. Therefore, their correlation is high. Negative extremes can be explained by one of the belt weighers not functioning. Here, values close to zero are paired with normal weight flows on the other belt weigher, leading to low r values.

Concluding the belt weigher data exploration, it can be said that the interplay of belt weigher AA101 and AA106 should not hamper model building in later stages of this research. With a proper pre-processing, no inconsistency can be found, and correlation is reasonably high. Problems occurred due to situations where one of the belt weighers detected a mass flow but the other not. These cases can be excluded by only including data points that lay above a mass flow of 0.3 [t/h].

The described insights deliver valuable information to the answering of the first sub research questions, as correlations and relationships in the belt weigher data were researched.

5.1.2 NIR-scanner Data – Area Flow

To explore the area flow of the NIR-scanners PO75 and PO75C, a joint distribution plot was created. For this, all area flows per scanner were summed up, to gather a first understanding of the overall behaviour. Results can be found in Figure 12.

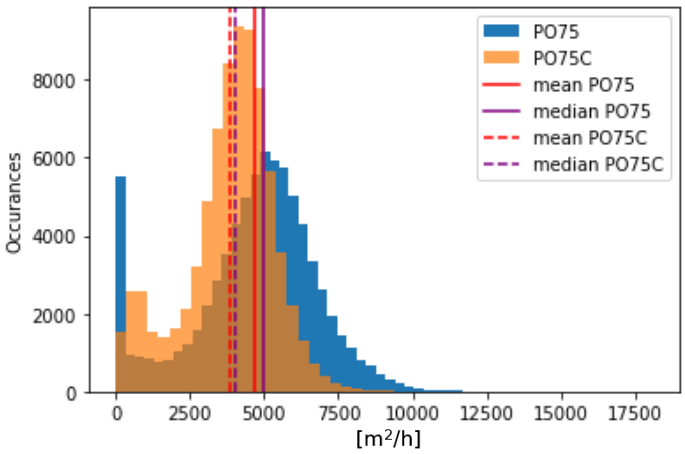

Figure 12: Distribution of occurrences of total area flows for PO75 and PO75C, using 50 bins for accumulation. Zeros were removed upfront.

Similar to the belt weighers, zeros were removed upfront and a left tail in the distribution can be observed. Furthermore, a shift of the left tail from PO75 to PO75C is seen. This raises the suspicion that PO75 does not measure 0 [m^2/h], when the belt is empty, but reports values close to zero instead. Through visual inspection of a distribution plot, with increased bin number, this suspicion could be confirmed.

In comparison to PO75, the tail of PO75C is right shifted. This can be due to the fact that PO75C is a cleaner. As the cleaning step aims to remove crucial impurities, a higher sensitivity of the scanner is expected. Therefore, if the belt is empty but PO75C still detects material, as result of a measurement error, the faulty data is probably of higher magnitude.

Following up with a more parameter driven analysis, the mean reduces from 4672.2 [m^2/h] to 3836.9 [m^2/h] from PO75 to PO75C. For both scanners, the median is slightly higher with values of 4955.1 [m^2/h] and 4058 [m^2/h]. Compared to the mass flow decrease of the belt weighers, the change in area flows feels rather low. This can be explained by the nature of the targeted material. At this stage of the separation process, PP and PE are mainly present as films and foils. Therefore, their area to weight ratio is fairly low. Furthermore, PP and PE represent the majority of the transferred material from PO75 to PO75C. Accordingly, the smaller magnitude of change, for the area flow compared to the mass flow, should result from a small area to weight ration of the discussed materials. Apart from the change of mean and median, it can be seen that the histogram extends up to 17,500 [m^2/h]. This represents some barely visible outliers on the higher end of the total area flows.

Delving into the material-based exploration, material-specific distributions plots were created. It was focused on PP_Film and BC as two examples that showed expected behaviour. Furthermore, PP and OTHER_POLYMERS were scrutinized, as they represent two examples that led to unexpected findings. The regarding plots are presented in Figure 13 and 14. Plots of all material-specific distributions can be found in Appendix 1.

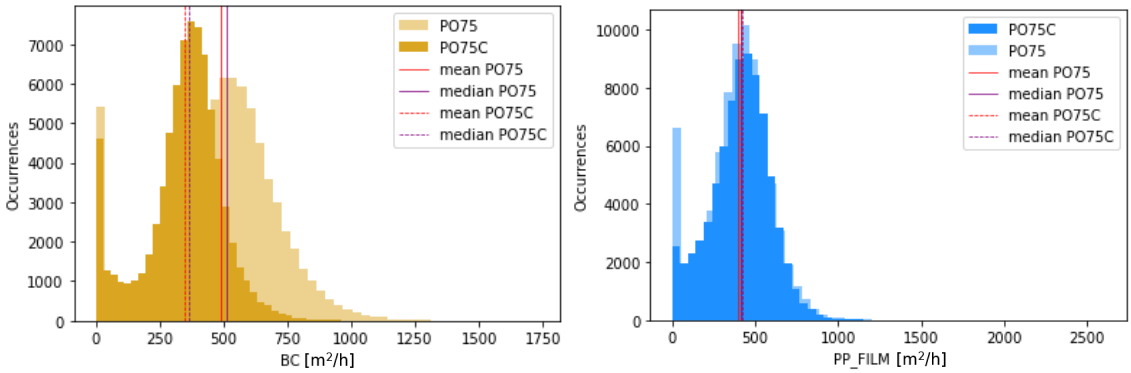

Figure 13: Material-specific area flow distribution for BC and PP_Film with zeros removed upfront.

It can be observed that the left tail of the distribution prevails in the material-specific plots. Furthermore, a greater magnitude of the tail for PO75 is present. Apart from that, both materials show the expected sorting behaviour. PP_FILM is expected to be fully transferred from rougher to cleaner, which is resembled by the overlapping distributions. For BC a decrease in material is presumed, as it resembles a non-targeted material. This is confirmed by the left shifted peak of PO75C as well as the left shifted mean and median.

Returning to the right-shifted tail of PO75C in Figure 12, another potential explanation emerges. The right shifted tail could be explained by the left shifted peaks for the separated materials on PO75C. Here, BC can resemble an example. Especially, materials with a lower area flow magnitude have their peak shifted to the left. Through this shift in magnitude and different separation efficiencies per material, the right shifted tail could occur. Materials that show the right shifted tail are PE_RIGID, PE_FILM, PET_BOTTLE and PET_MONO_TRAY. The respective plots are depicted in Appendix 1.

Going on to materials that did not show expected behaviour, the material-specific distributions for PP and OTHER_POLYMERS are presented in Figure 14.

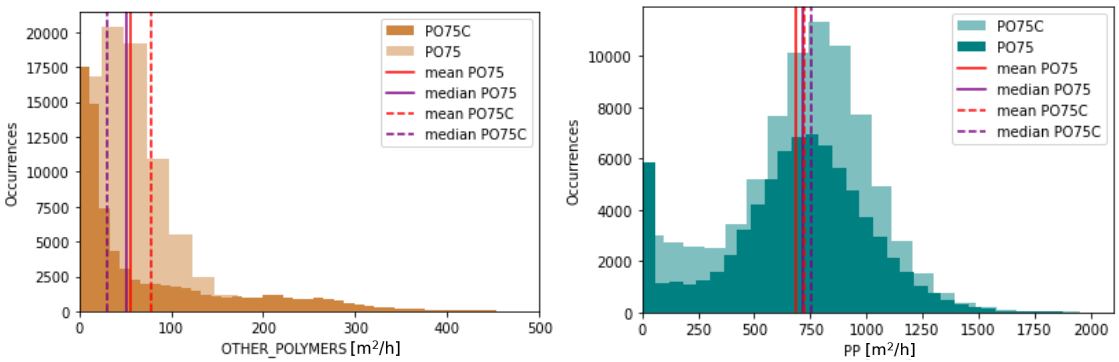

Figure 14: Material-specific area flow distribution for OTHER_POLYMERS (original x-axis range: 0 to 1300) and PP (original x-axis range: 0 to 5000), with zeros removed from the data beforehand.

On a first glance, the OTHER_POLYMERS plot passes the sanity check, as material occurrence seems to decreased from PO75 to PO75C. Nevertheless, observing the mean of the data, it is shown that the mean actually increases from 56 [m^2/h] to 78.3 [m^2/h]. As both distributions have the same amount of observations, an increasing mean indicates a greater detected total area. Due to the fact that from PO75 to PO75C material can only be removed but not be added, this change becomes impossible. Therefore, this finding can only be explained by a machine malfunction or different measurement behaviour of the scanners.

Different measurement behaviour of the scanners can be due to the inner workings of the unified material statistic, provided by TOMRA. This means that the statistic could compose the OTHER_POLYMERS category differently from scanner to scanner. Another reason could be the sensitivity regarding targeted and non-targeted materials. If there is a high priority that all particles of the targeted materials are detected, the machine could lean towards overclassifying these categories. This would mean that particles that do not belong to this group could be classified as targeted materials out of precaution. As both scanners have different target materials, this mechanism could overlap, which would amplify the described effect. Therefore, if the effect is strong enough, it could be the reason for seeing more material on the cleaner than initially detected on the rougher. The same explanation could be valid for the PP plot. Here, an increase of the mean from 686.8 [m^2/h] to 724.8 [m^2/h] is observed.

For the application of the modelling approaches, the correlation of the two scanners is of interest. Therefore, a scatter plot for visual inspection of the correlation between PO75 and PO75C was plotted. Results are depicted in Figure 15. Furthermore, r was calculated and is indicated within the plot.

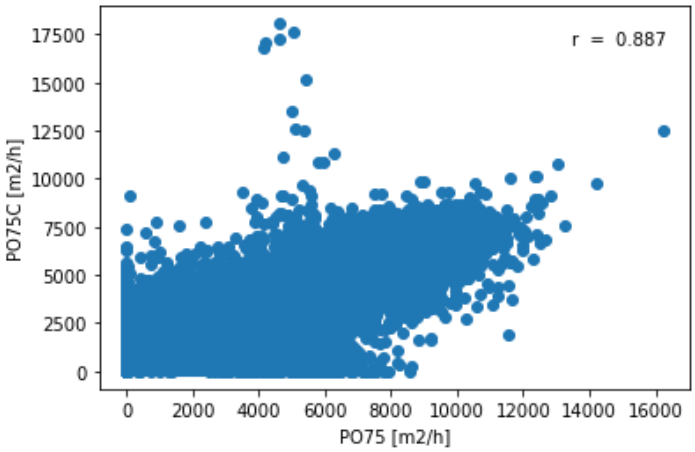

Figure 15: Scatter plots between NIR-scanner PO75 and NIR-scanner PO75C. The pearson correlation coefficient is indicated at the top left corner of the plots.

The Pearson correlation coefficient shows a strong positive correlation with a value of 0.887. Apart from that, a group of outliers can be observed for area flows around 5000 [m^2/h] for PO75. This could be a measurement mistake, as the outlier values triple the measured material from rougher to cleaner. As an alternative explanation, it could be imagined that particles got stuck between the two scanners. The release of the hold-back material later in time could lead to the elevated values on PO75C.

What furthermore sticks out is the high occurrence of measured material on one scanner, while the other scanner is empty or detects values close to zero. As empty entries for both scanners were removed, the hypothesis is made that the scanners occasionally measure values close to zero instead of zero. Therefore, these values were not removed upfront and the inconsistencies become tangible in the described plot.

Another possibility are temporal patterns in the data. Therefore, datapoints were ordered by time, split into 200 bins and for each bin r was computed. The result can be observed in Figure 16.

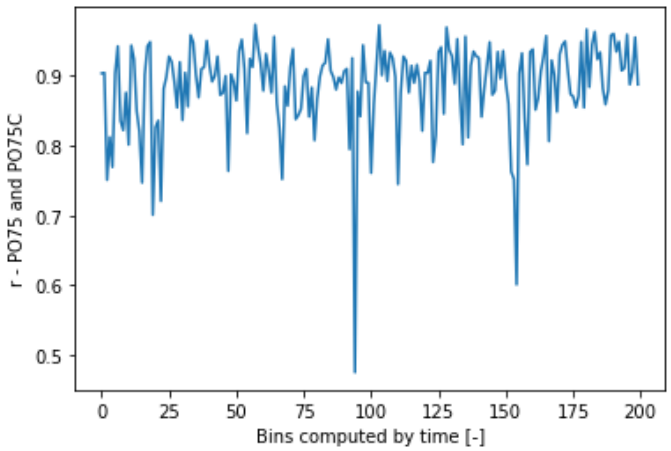

Figure 16: Pearson correlation coefficient for NIR-scanner PO75 with regards to NIR-scanner PO75C with 200 bins compiled by time.

Maximum correlation is indicated with 0.97, while minimum correlation was determined with 0.48. Except for two negative outliers, no temporal pattern can be observed. These outliers can be due to measurement errors. Furthermore, they could be based on the described suspicion that the scanners occasionally do not detect materials, but indicate values above zero. Apart from that, correlations toward the end of the plot can lead towards the hypothesis of a positive trend in the data. Nevertheless, if the plot would be truncated at bin number 40, the positive trend from bin 25 onwards could lead to the same hypothesis. Therefore, these kinds of trend detection should be done with caution. Moving on, an oscillation around the mean value of 0.89 can be observed, with amplitudes in the region of ± 0.1. This behaviour can be explained through the natural variance in the data and is therefore expected.

Continuing with the material-specific exploration between PO75 and PO75C, PP_FILM and BC were chosen as examples with expected behaviour and PVC and OTHER_POLYMERS as examples with unexpected behaviour. The regarding scatter plots are presented in Figure 17 and 18, while scatter plots for all materials are made available in Appendix 2. In general, PE_FILM and PE_RIGID were the best correlating materials with r values of 0.9 and 0.89, while OTER_POLYMERS and PS were the materials with the lowest correlations with r values of 0.07 and 0.13.

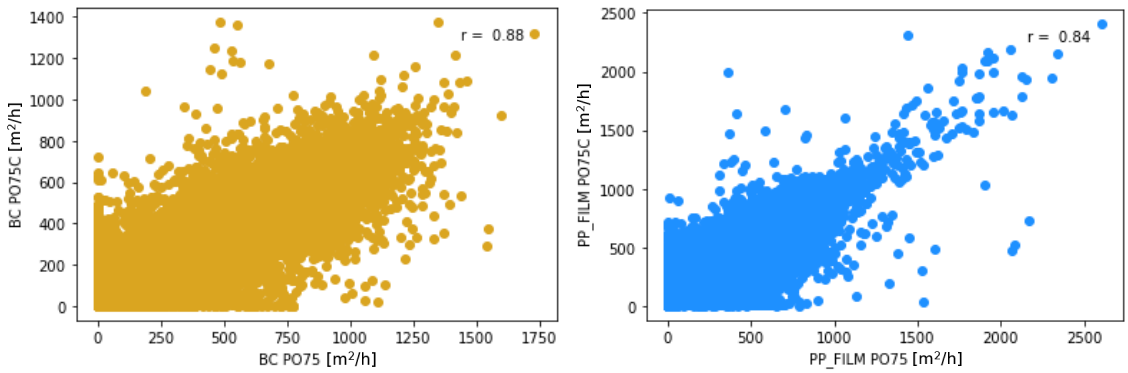

Figure 17: Correlations for BC and PP_FILM in form of scatter plots. The Pearson correlation coefficient is computed in the top right corner.

For BC and PP_FILM, high correlation values of 0.88 and 0.84 are reached. Outliers are present, but do not enter extreme value ranges. These positive circumstances indicate that both machines can identify the respective materials equally well and that the separation behaviour is as expected. Materials that show similar behaviour are PE_FILM, PE_RIGID, PET_MONO_TRAY, PET_BOTTLE and PP. PAPER performed reasonably well, but with a r value of 0.65 it has significantly lower correlation than the better performing materials.

Looking into the correlations for OTHER_POLYMERS and PVC this impression changes. Corresponding plots are depicted in Figure 18.

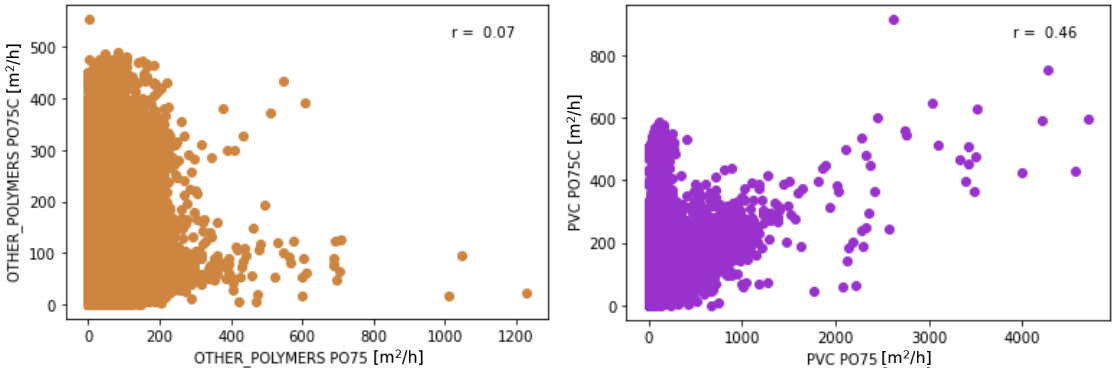

Figure 18: Correlations for OTHER_POLYMERS and PVC in form of scatter plots. The Pearson correlation coefficient is computed in the top right corner.

For OTHER_POLYMERS and PVC, low r values of 0.07 and 0.46 are observed. Ostentatiously, a majority of the values from PO75 get surpassed by values of PO75C. This is represented by the high accumulation of datapoints close to the y-axis. The presence of these datapoints is impossible, as material can only be removed but not added from PO75 to PO75C. This could be due to three reasons. First, PVC is a targeted material by PO75C but not by PO75. Therefore, the machine could be prone to over detecting this category. Second, the NIR spectra detected by the scanner is not always unambiguous to decide which material is present on the belt. Accordingly, each scanner follows a set of rules to decide the categorization. This set of rules is different for each machine. Reasons for ambiguous situations can be mixed materials or particles that lay too close to each other. Due to this vicinity, they can be confused as one bigger particle, instead of a number of small ones. When smaller particles are detected together as one larger particle, the machine still has to classify the entire particle. Therefore, if the smaller particles consist of different materials, inconsistencies arise and information about material is lost. Lastly, changing particle orientation on the belt can lead to different area flows on different scanners. NIR-scanners detect the area of a particle that is horizontally oriented to the belt. In other words, if a different side of the particle lays face up from one scanner to the next, the detected area changes.

Moving on, other not as well performing materials were EPS and PS. Comparing the good and the bad performing materials, it stood out that lower correlations occur with materials that have a lower magnitude area flow. Although this idea was not further tested, it could be an interesting starting point for further research.

Due to the materials with lower correlations, the idea emerged that there could be temporal patterns in the data. Therefore, the data was ordered by time, split into 200 bins and r was computed for each of them. Again, two well performing materials and two poorly performing materials were selected. For the good performing materials, BC and PP_FILM were chosen while OTHER_POLYMERS and PVC were picked respectively. The plots can be found in Figure 19 and 21. Plots for all materials are presented in Appendix 3.

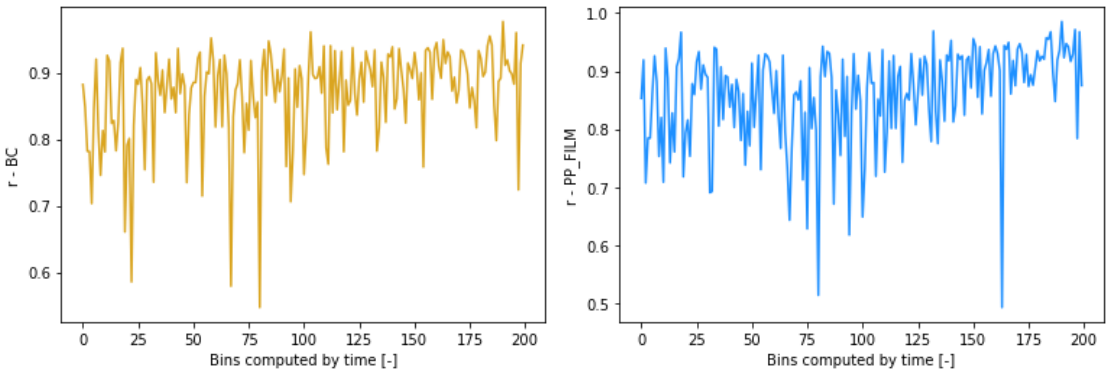

Figure 19: Pearson correlation coefficient for BC and PP_FILM regarding NIR-scanner PO75 and PO75C with 200 bins compiled by time.

No clear patterns can be observed for BC or PP_FILM. A slight positive trend is suspected towards the later bins, but is most probably due to normal fluctuation. Apart from that, a few negative outliers are seen. Interestingly these outliers do not overlap. This could be due to material-specific machine malfunctions. An alternative explanation is that the input material varied exceptionally from the norm for the specific materials. This could lead to an altered separation behaviour, having changed correlation as consequence. Examining the frequency of outliers, with only three outliers for BC and two outliers for PP_FILM their occurrence seems acceptable. The highest correlation is 0.98 for BC and 0.99 for PP_FILM, while the lowest correlation is 0.55 and 0.49 respectively. Scatter plots for the bins with the lowest r value were computed to generate an idea for the reasons of the outliers. The results can be found in Figure 20. Plots for the lowest correlating bins of all materials are presented in Appendix 4.

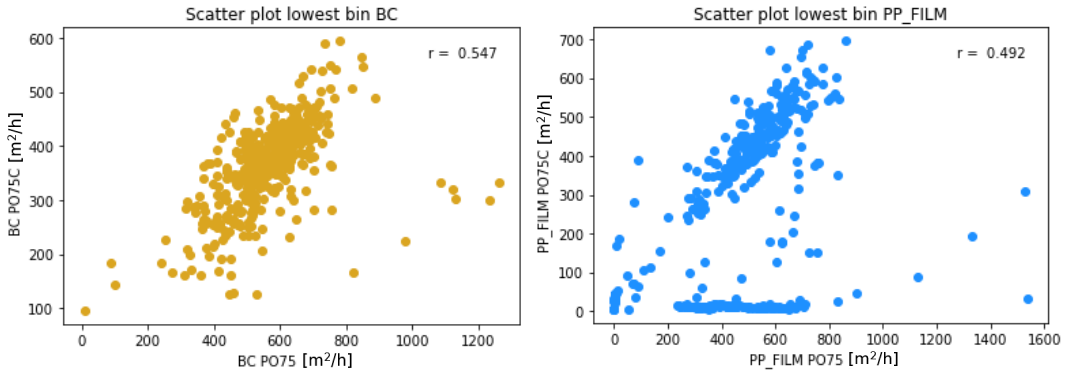

Figure 20: Correlations for BC and PP_FILM in form of scatter plots for the lowest performing correlation bin by time. The Pearson correlation coefficient is indicated at the top right corner of the plot.

Starting off with BC, the scatter plot does not reveal any new inconsistencies. Correlation is lowered by some outliers around the point cloud. These outliers depict higher values for PO75 than PO75C, which is physically possible. This could indicate an especially good separation for these occurrences, as less material then normally reached the next scanner. As BC is supposed to be dropped, this symbolizes an expected behaviour.

Delving into the reason for the lowered correlation, the idea of a connection to the area flow magnitude on the separating scanner was risen. Therefore, the mean of the lowest performing bin was compared to the mean of the complete data set for PO75. Here, an increased mean of 559.96 [m^2/h], compared

to 489.2 [m^2/h] for the complete data set was determined for BC. This can be explained by the outliers on the right-hand side of the plot. Here, either especially large particles or a high occurrence of normal particles was successfully separated. As a result, the correlation was lowered and the mean increased for BC on PO75.

Focusing on PP_FILM, the suspicion of events where one scanner detects material why the other scanner does not detect material gets substantiated. Another explanation for the accumulation of data points close to zero for PO75C could be a temporal malfunction of the separation mechanism. PP_FILM is targeted by PO75. Therefore, a malfunction of the air nozzles could explain the occurrences where PO75 detects the material but it does not reach PO75C. Comparing the mean of the complete data vs. the data of the lowest performing bin, an increased mean of 457.9 [m^2/h] compared to 399.5 [m^2/h] is observed. In this case, the explanation could be the other way around. Too much material could have been present on the belt hindering the ejection of PP_FILM. This could also explain the occurrences where PP_FILM was detected on PO75 but not on PO75C.

Examining the lower performing materials, the correlations of OTHER_POLYMERS and PVC for 200 bins computed by time are depicted in Figure 21.

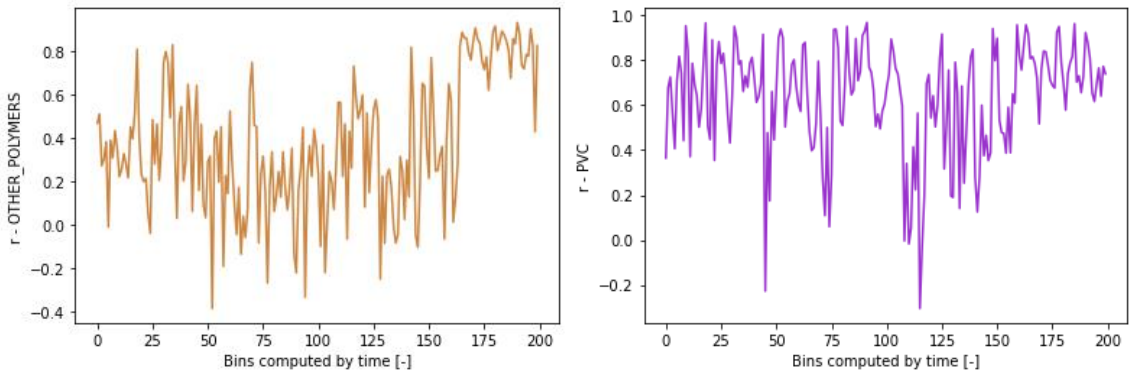

Figure 21: Pearson correlation coefficient for OTHER_POLYMERS and PVC regarding NIR-scanner PO75 and PO75C with 200 bins compiled by time.

For OTHER_POLYMERS, an oscillation of the correlation around 0.3 with an approximate magnitude of +0.3 and -0.4 is observed. Furthermore, a positive shift towards the end of the data is seen. This shift could be due to a change of programming of the machine or some other special event. For PVC, no clear trends are observed but two outliers are visible in the plot. These outliers do not overlap with the outliers of Figure 19. Therefore, material-specific malfunctions or special events could be the explanation. An example for a special event could be a heavy material occurrence, which hampered separation. The highest correlation is seen with 0.93 for OTHER_POLYMERS and with 0.97 for PVC, while the lowest correlation is detected with -0.38 and -0.3 respectively. To gain better insight into the least correlating parts of the data, scatter plots for the bins with the lowest correlation were plotted. The scatter plots for OTHER_POLYMERS and PVC and are presented in Figure 22.

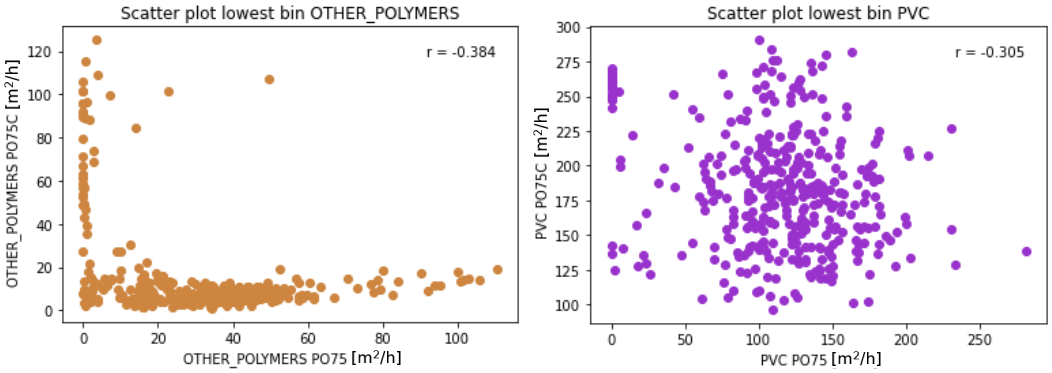

Figure 22: Correlations for OTHER_POLYMERS and PVC in form of scatter plots for the lowest performing correlation bin by time. The Pearson correlation coefficient is indicated at the top right corner of the plot.

For OTHER_POLYMERS, a clear trend is visible. High values on PO75 paired with low values on PO75C indicate a successful separation. Apart from that, the high occurrence of values that are close to zero for PO75 paired with values not close to zero for PO75C explains the negative correlation. As the addition of material is impossible, a malfunction of OTHER_POLYMER detection on PO75 is suspected for these instances. Scrutinizing the mean of the observed bin compared to the mean of the complete data set, an increase from 31.5 [m^2/h] to 56 [m^2/h] is observed. Resulting of this, another potential explanation for impossible material increases arises. Increased area flow means could indicate situations where the mono layer of the material on the conveyor is not given. This can lead to overlapping particles. Therefore, particles that have another particle on top cannot be detected. If in the overlapped material OTHER_POLYMERS is present in high volumes, the impossible values can be explained.

Regarding the PVC scatter plot, no clear pattern is seen. Examining the scaling of the x- and the y-axis, one can observe that the majority of the values lies in the impossible range. This means that values from PO75C have greater magnitudes then values from PO75. As PVC is meant to be ejected on PO75C, the sensitivity of the classification could be the reason for this occurrence. Comparing the mean of the lowest performing bin to the complete data set, a decrease from 132.5 [m^2/h] to 110.9 [m^2/h] is seen. An explanation for the joint occurrence of low correlation and a decreased mean could be that with less material present on the belt inconsistencies or special events have a higher influence on the correlation. In these cases, the inconsistencies represent a bigger percentual share of the total material, wherefore their impact is higher. Nevertheless, this is contradicted by the upfront created hypotheses for the joint occurrence of low correlations and increased means. Therefore, an analysis of which effect is dominant in which situation could be of great interest.

To give this analysis a start, plots for the lowest correlation bin with total area flows from PO75 and PO75C as well as material-specific area flows for PO75 and PO75C were compiled. Here, it was focused on two well performing materials and two low performing materials. The materials are BC and PP_FILM as well as OTHER_POLYMERS and PVC. The created plots are depicted in Figure 23 and 24. Plots for all materials are presented in Appendix 5.

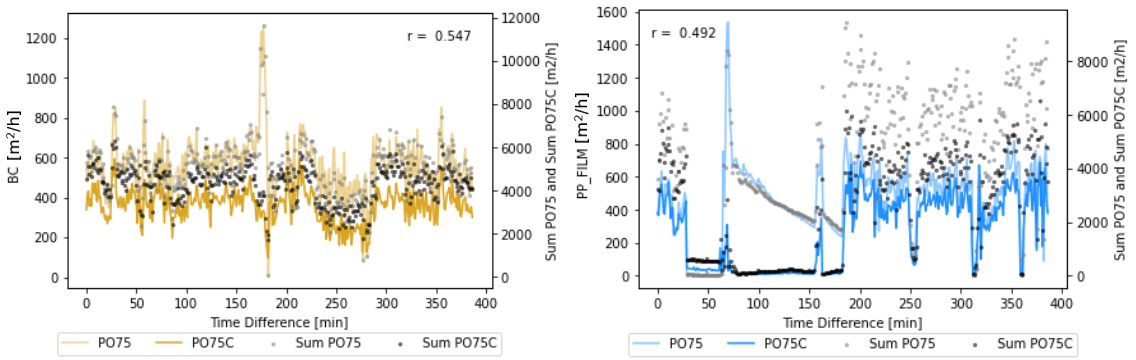

Figure 23: Total area flows on PO75 (grey dots, right axis) and PO75C (black dots, right axis) as well BC and PP_FILM for PO75 (light lines, left axis) and PO75C (solid lines, left axis).

In the Figure, black dots represent total area flows on PO75 and grey dots indicate total area flows for PO75C. Their magnitude is indicated on the right y-axis. The dark line shows the values for BC on PO75C and the light line indicates BC on PO75. Both lines refer to the left y-axis. It is seen that BC performs as expected. The light line stays on top of the dark line for the majority of the time steps. Also, the total area flows act accordingly. This means that black dots are present below grey dots for most of the observed datapoints. A noticeable exception is represented by the peak of PO75 and the depression of PO75C close to bin 170. On PO75, an accumulation of specific and total material is happening. At the same time, the total and specific material decreases on PO75C. Right after, PO75C jumps back to normal behaviour, while PO75 drops to zero for the total and specific area flows. This could be due to an over occupation of PO75. An over occupation on PO75 potentially leads to decreased separation successes, wherefore less material in total reaches PO75C. To explain the drop to zero of PO75, it is imaginable, that as a reaction to the over occupation the conveyer belt was emptied. The less extreme drop for PO75C could be due to the time delay between the scanners. This means that the phase where the belt is empty could be spread over two time steps for PO75C. Therefore, no drop to zero but a drop close to it is detected.

Way more questions than for BC arise while examining the plot for PP_FILM. The first observation that can be made is around bin number 50. Here, PO75 drops close to zero, while PP_FILM for PO75C stagnates around 100 [m^2/h] and the total area flow of PO75C close to 500 [m^2/h]. This stagnation over a course of roughly 35 bins is unusual compared to the fluctuation for the rest of the data. Therefore, a constant offset for special cases during measurement can be suspected. Directly after this, a maximum of roughly 9000 [m^2/h] occurs for PO75. At the same time PO75C detects area flows close to zero. This peak could originate from stuck material that was released in one go. Possibly this is an explanation for the empty belt upfront, as the accumulated material could resembles the material that did not reach the scanner before.

The difference in detection from PO75 to PO75C could be due to a measuring error. This is suspected as PP_FILM is a target material of PO75. Therefore, it should be transferred to PO75C, but PO75C stays close to empty. An explanation for this could arise from the comparison of summed and material-specific area flows on PO75. It can be observed that values decline in a linear fashion with a similar slope. Checking the data for other materials, in the same time period, a similar behaviour can be observed for BOARD_CT, PET_BOTTLE, PET_MONO_TRAY and PP. This could be due to a stopped belt or to material that got stuck on the detection area of the scanner. The non-moving material could

then be successively registered as background over time, what would explain the linear decline of detection.

In Figure 24, plots for the joint analysis of summed and specific area flows for PO75 and PO75C regarding OTHER_POLYMERS and PVC are depicted.

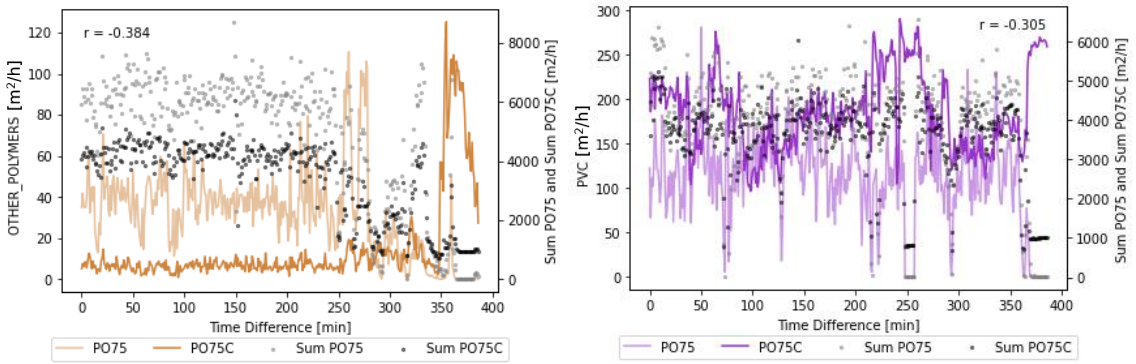

Figure 24: Total area flows on PO75 (grey dots) and PO75C (black dots) as well OTHER_POLYMERS and PVC for PO75 (light lines) and PO75C (solid lines).

A lower correlation can be observed for OTHER_POLYMERS and PVC. The OTHER_POLYMERS plot shows greater magnitude of variation for PO75 than for PO75C. This lowers the correlation, as for greater change on PO75 lower change on PO75C is seen. Apart from that, from bin 280 onwards, the material-specific lines start to swap position. This symbolizes a switch from more OTHER_POLYMERS on PO75 to more OTHER_POLYMERS on PO75C. As this is impossible, a measurement error needs to be at hand. For the occurrences around bin 290 and 320, the area flow sum of PO75 drops close to zero, wherefore a malfunction can be assumed. From bin 350 onwards, OTHER_POLYMERS peaks heavily on PO75C and starts to decline again. With a time delay, this peak is mirrored on PO75 and then converts into a stagnation for both sums of the scanners. Here, P075 stagnates close to 0 [m^2/h], while the stagnation for PO75C is observed around 1000 [m^2/h]. It is interesting that the same stagnation pattern as for PP_FILM can be observed. This raises the suspicion, that when the sum of PO75 is close to zero, PO75C still detects a sum of roughly 1000 [m^2/h]. Another hunch that emerges is that OTHER_POLYMERS somehow could balance other material categories during stagnation, as it declines, while the sum of the scanner stays unnaturally stable.

For PVC, the material-specific area flows measured on PO75 and PO75C develop in different directions on a frequent basis. Strong examples resemble opposite peaks around bin 70 and the area between bin 220 to 240. Furthermore, the encountered stagnation pattern can be also observed for PVC arounf bin 250 and from bin 370 onwards. Interestingly, a third behaviour for this case can be seen in the data. PP_FILM dropped with the sum of the scanners and stagnated at a low value, OTHER_POLYMERS peaked and declined during stagnation, while PVC peaks during stagnation and oscillates at high values. These behaviours could not be reproduced in the time periods of lowest correlation bins of other materials, but they reveal a range of behaviours the scanners show during these stagnations. This leads to the suspicion that a state of the scanner exists where it tries to keep the sum of detected area flows constant. In these cases, some materials deliver a constant contribution, while other materials balance each other out through declining or increasing over time.

The described explorations motivate several pre-processing procedures that will be tested later in this work for model building. Through the depicted histograms and the found stagnation for the lowest correlation bins, upper and lower area sum thresholds were considered good candidates for data pre-processing. In general, if a data point gets removed for one scanner it is also removed for the other scanner to avoid NaN values and resulting calculation errors. To test the thresholds out, it was decided to remove all data points that have a lower total area flow of 1000 [m^2/h] for PO75 and a lower total area flow than 1250 [m^2/h] for PO75C. This was done due to the discovered stagnations around 1000 [m^2/h] for PO75C and the left tail of the distribution for both scanners in the initially discussed histograms. These histograms also were the motivation to implement the upper threshold to remove outliers from the data. Here, an upper threshold of 12,500 [m^2/h] and 10,000 [m^2/h] was selected for PO75 and PO75C. A histogram with the applied thresholds for total area flow on PO75 and PO75C can be found in Figure 25.

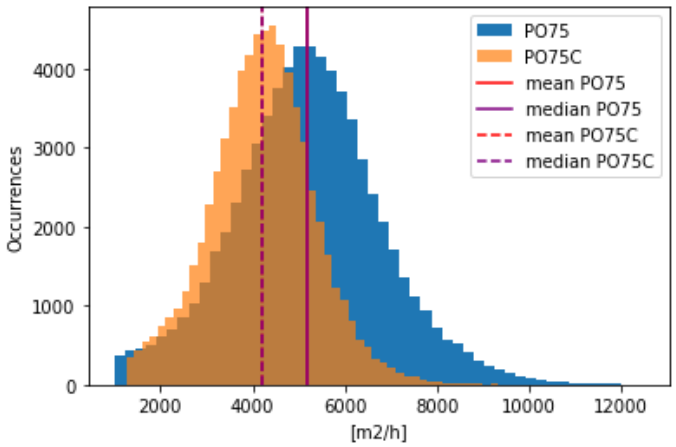

Figure 25: Distribution of occurrences of total area flows for PO75 and PO75C, using 50 bins for accumulation. Zeros were removed upfront and thresholds of 1000 and 12500 as well as 1250 and 10000 were applied for PO75 and PO75C respectively.

Applying the thresholds results in normal distributed data. Furthermore, a slight truncation at the left side of the distribution and centred means and medians for both scanners are observed.

Concluding the NIR-scanner area flow data exploration, it can be said that the data should be used with care. Especially, for occurrences of low magnitude total area flows or low correlations, inner workings of the NIR-scanners were encountered that are not understood. Furthermore, attention should be given to occasions where material that is detected on PO75C exceeds material detected on PO75. These occurrences are impossible and can only be explained to a limited amount by belt over occupations, stuck material that gets released in bulk or fluctuations due to changed particle orientation. Furthermore, situations where one NIR-scanner is detecting material, but the other NIR is detecting no material at all need to be considered. Interestingly enough these events were present on a material-specific level and for the total area flows.

The obtained information provides valuable insights to answer the first sub-research question. For this, correlations and relationships in the data have to be analysed. This was successfully done for NIR-scanner area flow information.

5.1.3 NIR-scanner Data – Belt Occupation

Position specific NIR-scanner data regarding belt occupation was made available by TOMRA. For this, the belt is split up into 70 pieces over its width. While the machine is running, the total material occurrence is measured for each of these sections. The analysis of this data will help solving sub research question one. An exemplary graphical representation of the belt occupation information can be found in Figure 26. The Figure resembles a screenshot from the TOMRA Insight portal.

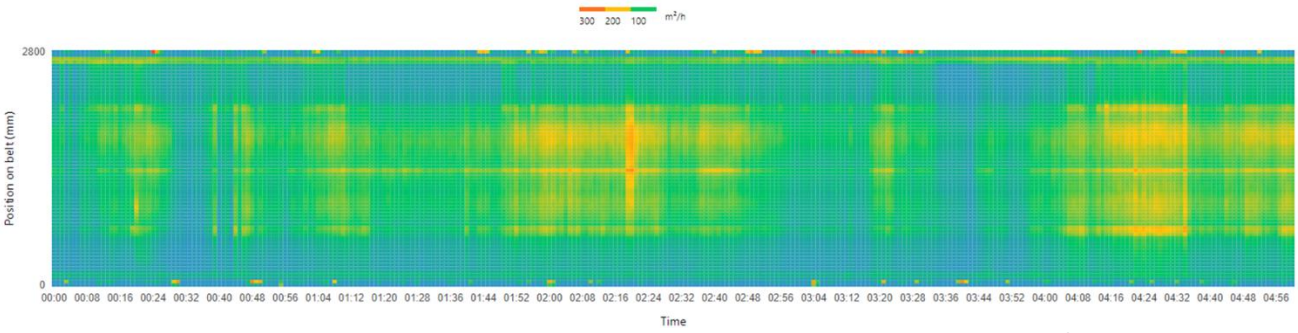

Figure 26: Exemplary graphical representation for the specific belt occupation in [m²/h] provided by TOMRA. (TOMRA, 2024)

To make the data usable for model building and data exploration, the provided information was aggregated. For this, the occurrences of belt occupations were counted between 0-50 [m^2/h], 50-100 [m^2/h], 100-200 [m^2/h], 200-300 [m^2/h] and >300 [m^2/h]. The used steps were inspired by the color-coding scheme used by TOMRA. For this, information displayed for PO75 and PO75C, in the TOMRA Insight portal, was scrutinized and the categories were derived. In a first step, the distributions of the different categories were plotted and are presented in Figure 27.

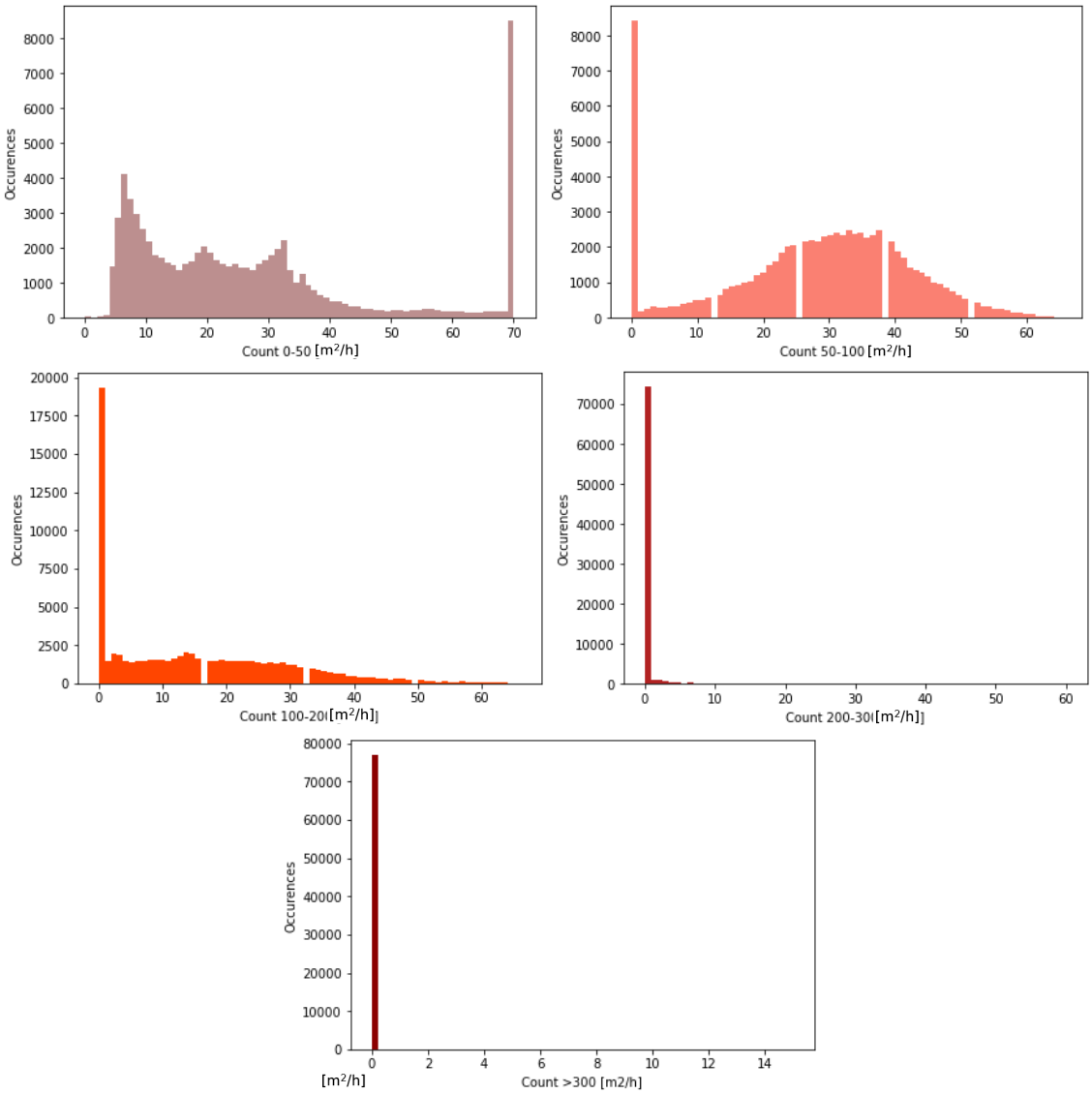

Figure 27: Distributions in form of histograms for counted belt occupancies on PO75.

The 50-100 [m^2/h] category has the greatest number of total occurrences with more than 2.14 million counts. Directly after, the 0-50 [m^2/h] category follows with 2.08 million counts. On the lower end is the 100-200 [m^2/h] category with 1.17 million counts, while the 200-300 [m^2/h] category and the >300 [m^2/h] are rarely registered with 11,704 and 1,144 occurrences.

The only range of belt occupations that is visibly normal distributed, except for its left tail, is 50-100 [m^2/h]. All other categories do not follow a common distribution, based on visual inspection. What can be noticed is that 0-50 [m^2/h] has a strong right tail, while all other categories have a strong left tail. This means that there is a part of the data where all sections are not over occupied. This is resembled by 70 counts for the 0-50 [m^2/h] category for over 8500 occurrences. These 8500 occurrences are mirrored by the other categories having a count of 0 occurrences. For 50-100 [m^2/h], the number of 0 counts does not exceed the number of 70 counts for 0-50 [m^2/h]. Therefore, when not all sections are within the 0-50 [m2/h] category, always sections that are in the 0-50 [m^2/h] and in the 50-100 [m^2/h] category occur. For all remaining categories, the amount of 0 occurrences exceeds this value, wherefore they are less often present. The 100-200 [m^2/h] category has a maximum number of

66 counts, while the 200-300 [m^2/h] category and the >300 [m^2/h] category have maximum counts of 60 and 15. Furthermore, the 0 counts increase dramatically. For the 100-200 [m^2/h] category they lie at 19,316, for the 200-300 [m^2/h] category they already go up to 74,188 and for the >300 [m^2/h] category they increase to 77,055. These described occurrences and distribution are as expected. While very high to high occupations occur rather rarely, moderate to low occupancies represent the majority of the data. This is wanted for favourable separation conditions.

Starting the material-specific exploration, it was checked if occupation counts correlate with material occurrences. For this, scatter plots with the different materials and the occupation counts were compiled. Furthermore, for each material a version was added where the coloration of the data points is set to 0.5%. This means that for full correlation 200 data points or more have to overlap. This generates a better feeling for the density of the data. To not discuss all available materials, four examples were selected. BC and PP_FILM show distinguishable patterns, while OTHER_POLYMERS and PVC are harder to interpret, wherefore they have been picked. This was done to show the range in between the materials. The plots can be found in Figure 28 and 29 respectively. Scatter plots for all materials are presented in Appendix 6.

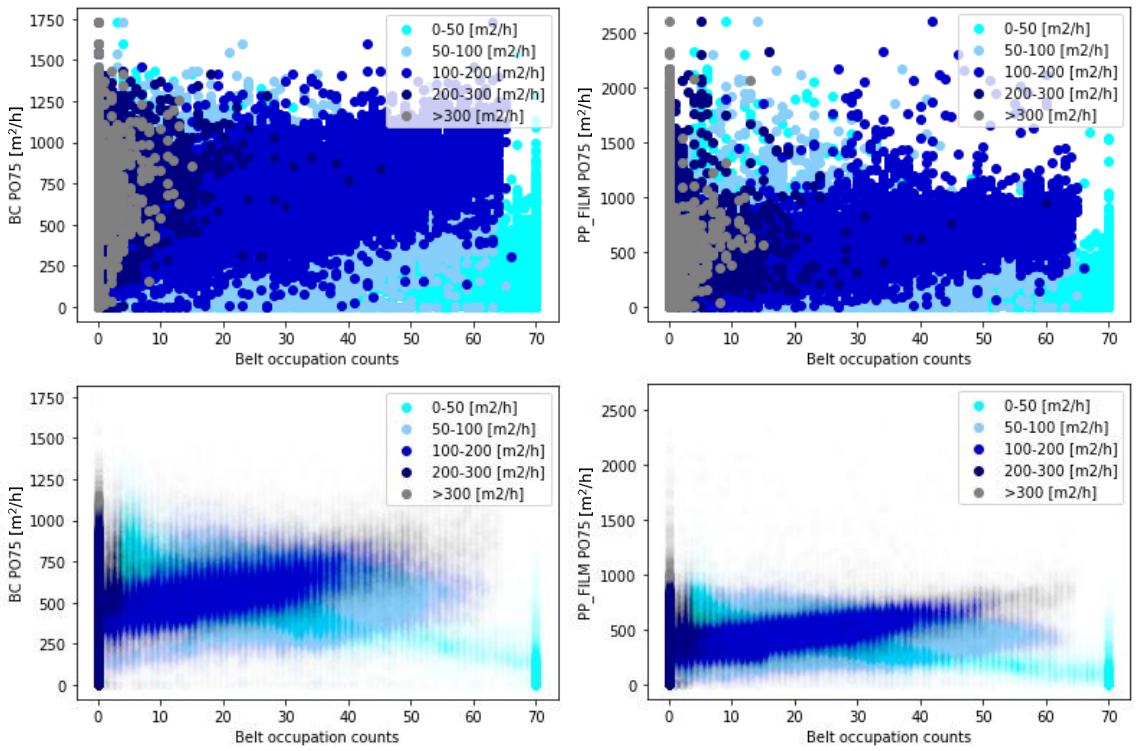

Figure 28: Scatter plots for BC and PP_FILM area flows on PO75 together with belt occupation counts from PO75. One version with full coloration (upper plots) and one version with 0.5% coloration (lower plots).

Both materials show rather similar behaviour. In the 100% coloration plots a slight difference is seen for the >300 [m^2/h] and 200-300 [m^2/h] category. For PP_FILM, higher occurrences of these categories are present in the lower half of the y-axis. For BC, they are placed towards the middle of the y-axis. At a first glance this seems counterintuitive. One would expect a high amount of the highest category occurrences on areas with high magnitudes area flows. Nevertheless, it is revealed that these two categories are not responsible for the high area flow magnitudes on a common basis. Instead, the 100-200 [m^2/h] category is governing. Therefore, it can be claimed that the >300 [m^2/h] and the

200-300 [m²/h] categories are more exceptions than useful indicators. In the 0.5% coloration plots, more patterns become visible. The 100-200 [m²/h] category shows a positive trend with high magnitude area flows. On the other hand, the 0-50 [m²/h] category shows a negative trend together with a balancing peak at the right side of the plot. Especially, the high area flows in this peak resemble an ideal case. This is because a high amount of material is spread out evenly enough to avoid over occupation, but still reaches the detected throughput. The 50-100 [m²/h] category acts as balancing ground noise. It has no positive or negative correlation and a normal distribution around 500 [m²/h] for both materials. This could not add any explanatory power to a linear model. Nevertheless, the peak around 60 counts for both materials could help to predict an area flow of 500 [m²/h]. This is valid for both materials, but would need to be leveraged with a non-linear or machine learning model.

Moving on to the correlations between belt occupation and OTHER_POLYMERS, as well as PVC, 0.5% coloration plots and full coloration plots are depicted in Figure 29.

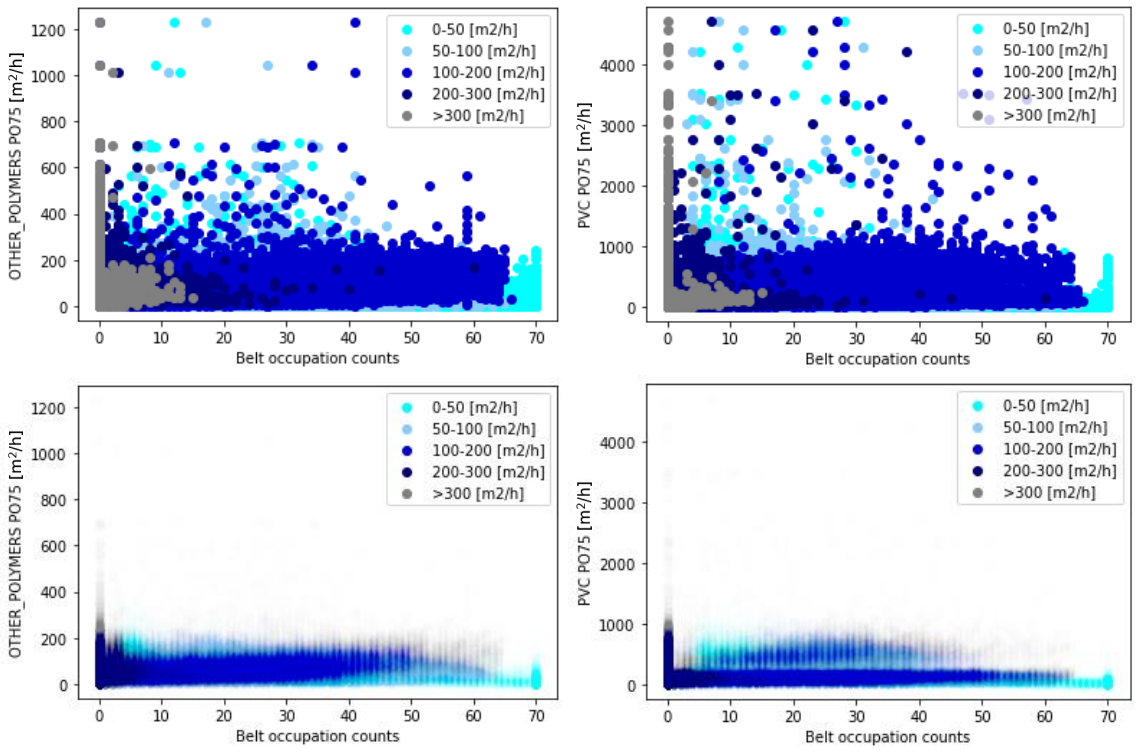

Figure 29: Scatter plots for OTHER_POLYMERS and PVC area flows on PO75 together with belt occupation counts from PO75. One version with full coloration (upper plots) and one version with 0.5% coloration (lower plots).

For the full coloration plots, both materials behave similarly. The 0-50 [m²/h] and 50-100 [m²/h] categories show equal occurrences of all counts for values up to 300 [m²/h] for OTHER_POLYMERS and for values up to 1000 [m²/h] for PVC. Afterwards, a rather random occurrence of categories, for values above these thresholds can be observed. Furthermore, for data points above 45 belt occupation counts, only the 100-200 [m²/h] category is depicted, for area flows above the thresholds. This observation could resemble the fact that after a certain threshold of occurrences, material sums can only be realized by categories that allow for enough area flow in combination with high enough belt occupation counts. This time the >300 category has their highest counts on the lower level of area flows. This could lead to the conclusion that with over occupied areas on the belt, PVC and

OTHER_POLYMERS do not get detected as well as normal. The same applies for the 200-300 [m^2/h] category. This strengthens the idea that the 100-200 [m^2/h] category is governing while the >300 [m^2/h] and 200-300 [m^2/h] categories are less influential due to their low occurrence. Furthermore, they could indicate hindered detection. The 0.5% coloration plots only reveal slightly more information. For OTHER_POLYMERS a vague positive trend of the 100-200 [m^2/h] category can be seen. The 0-50 [m^2/h] category has a right tail for both materials, but no trend is visible for the other parts of the data. Apart from that, other categories do not show any clear patterns. This means that these categories most probably do not have great influence on the appearance of the materials.

Another idea for the belt occupation data is, that for over occupied parts of the belt, separation could be hampered. Therefore, occupation counts were plotted against area flow data from PO75C and are presented in the form of scatter plots. BC and PP_FILM are used as positive examples, while OTHER_POLYMERS and PVC are showcased as negative examples. Plots can be found in Figure 30 and 31, while representations of all materials are depicted in Appendix 6.

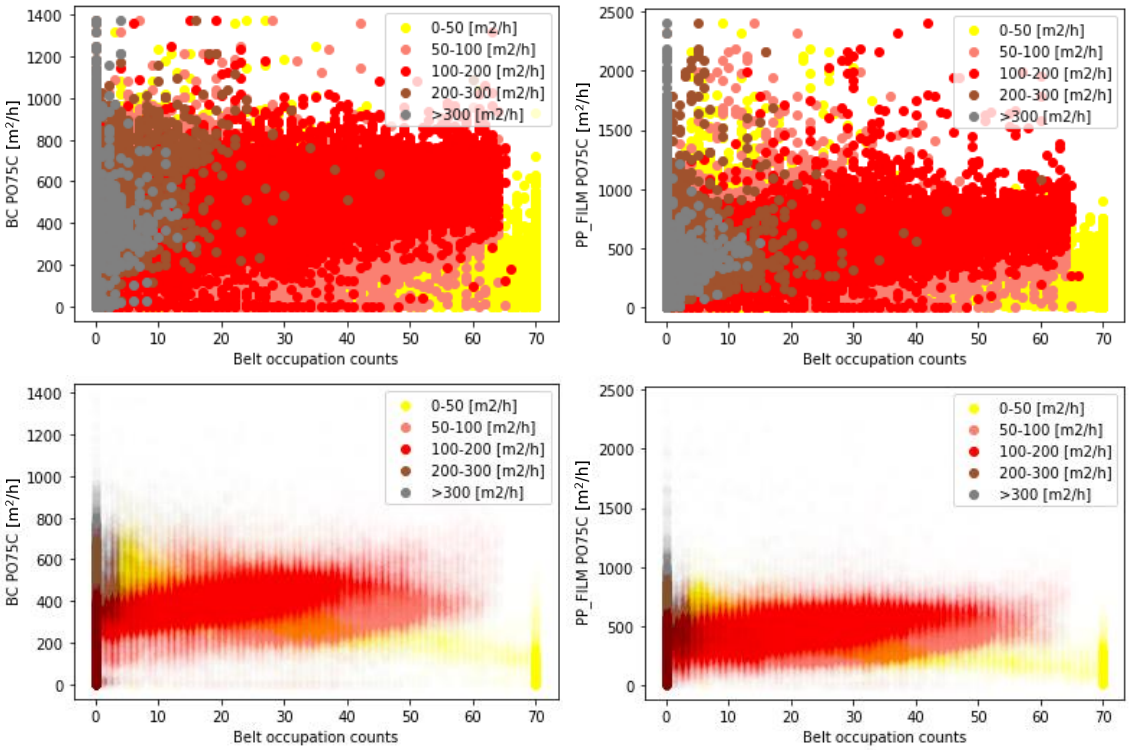

Figure 30: Scatter plots for BC and OTHER_POLYMERS area flows on PO75C together with belt occupation counts from PO75. One version with full coloration (upper plots) and one version with 0.5% coloration (lower plots).

In general, similar trends as in the plots for PO75 can be observed. A decreased area flow is seen for BC because of the separation from PO75 to PO75C. For PP_FILM, similar area flows occur due to the successful transference of target material. Considering the 50-100 [m^2/h] category, a positive trend could be suspected. Nevertheless, scrutinizing a scatter plot of only this category, it is revealed that the category performs as for PO75. The difference is that this category has a reduced occurrence for up to bin 15 for PP_FILM and BC, wherefore the joint plot appears differently. Furthermore, it can be observed that for BC the 200-300 [m^2/h] category seems to have more counts for higher area flows.

Nevertheless, in the 0.5% coloration plot it is revealed that the density of these values lies in the lower half of the area flow magnitudes.

Starting the analysis of the belt occupation on PO75 together with the area flows on PO75C for OTHER_POLYMERS and PVC, scatter plots were compiled. A 0.5% coloration and full coloration version for both materials are presented in Figure 31.

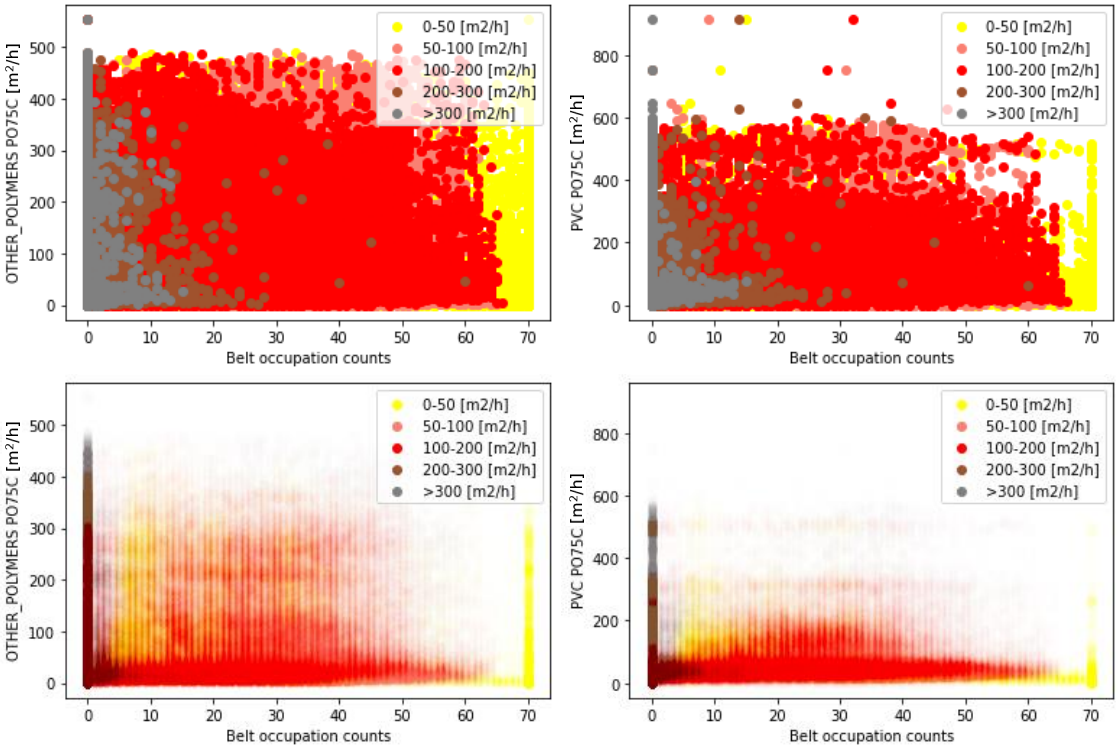

Figure 31: Scatter plots for OTHER_POLYMERS and PVC area flows on PO75C together with belt occupation counts from PO75. One version with full coloration (upper plots) and one version with 0.5% coloration (lower plots).

Scrutinizing the plots for OTHER_POLYMERS and PVC, similar behaviour to the PO75 plots can be observed. Nevertheless, compared to the 0.5% plot of PO75, in the PO75C version, no trend for OTHER_POLYMERS is visible. Furthermore, the data points in the full coloration plot are more clustered, without the scarce point cloud above the dense data point area, as in the PO75 versions. This could be due to the successful removal of high area flow outliers during separation. In contradiction to this hypothesis, for PO75, the dense data point area of OTHER_POLYMERS is seen up to 300 [m^2h], while for PO75C it goes until 500 [m^2/h]. For the 0.5% coloration plot, these values are 200 [m^2/h] and 300 [m^2/h] respectively. For PVC, this dense data point area decreases intensely from 1000 [m^2/h] to 600 [m^2/h] for PO75C. Additionally, a decrease from 800 [m^2/h] to 200 [m^2/h] can be observed for the 0.5% coloration plot. This is backed up by the mean of OTHER_POLYMERS, of 56 [m^2/h] for PO75, and 78.2 [m^2/h] for PO75C, as well as 132.5 [m^2/h] and 57.9 [m^2/h] for PVC. While this is an expected behaviour for PVC due to the sorting, OTHER_POLYMERS should normally decrease. Therefore, a changed classification could be an explanation.

Concluding, slight correlations between belt occupation and material flow magnitude can be seen. Expected relationships between PO75C and belt occupation could not be confirmed. This was initially suspected due to expected changes in sorting efficiency, triggered by altered belt occupations.

Nevertheless, similar correlations as for PO75 are observed, but with a decreased significance. While this information has limited use for linear model building, enough non-linear trends are observed to suspect added value in machine learning applications.

The gained knowledge completes the information that is needed to answer the first sub research question. All available data sources were explored and insights about correlations and relationships that are present in the data were gathered. Therefore, model building and prediction tasks can be conducted on a well-informed basis and are describe in the following subchapters.

5.2 Statistical Modelling – Area Density

To join belt weigher and NIR-scanner data, an OLS modelling approach was chosen. This was done to determine the area densities for the processed materials. Determination of area densities is a crucial, as it enables transference from area flow data into mass stream data, which necessary for reliable quality predictions. Quality prediction could be done in m^2, but due to the missing information about the height of the particles, this prediction results in high variance of final volume or mass. The outcomes of this subchapter will, therefore, deliver an important contribution to the answering of sub research question two.

5.2.1 OLS use case

Area densities multiplied by area flows result in mass flows. Therefore, summing up area flows that are multiplied with area densities, for all available materials of a NIR-scanner, result in the mass input of the same. This can be seen as a linear model describing the relationship between area flows, area densities and mass flows. A mathematical formulation of the described relationship can be found below.

$$\begin{bmatrix} \dot{m}_{t1} \\ \dot{m}_{t2} \\ \dot{m}_{t3} \\ \dot{m}_{t4} \\ \dot{m}_{t5} \\ \dots \end{bmatrix} = \begin{bmatrix} \dot{A}_{mat1,\ t1} & \dot{A}_{mat2,\ t1} & \dot{A}_{mat3,\ t1} & \dots \\ \dot{A}_{mat1,\ t2} & \dot{A}_{mat2,\ t2} & \dot{A}_{mat3,\ t2} & \dots \\ \dot{A}_{mat1,\ t3} & \dot{A}_{mat2,\ t3} & \dot{A}_{mat3,\ t3} & \dots \\ \dot{A}_{mat1,\ t4} & \dot{A}_{mat2,\ t4} & \dot{A}_{mat3,\ t4} & \dots \\ \dot{A}_{mat1,\ t5} & \dot{A}_{mat2,\ t5} & \dot{A}_{mat3,\ t5} & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix} * \begin{bmatrix} \rho_{mat1,\ predicted} \\ \rho_{mat2,\ \text{predicted}} \\ \rho_{mat3,\ predicted} \\ \dots \end{bmatrix} \tag{5.1}$$

As one can see, this formula also represents the form of an OLS model. Here, the mass flow act as dependent variable, while the area flows represent the independent variables and the area densities are the estimators. Applying the OLS model like this, it is not used in the classical way. As the goal is to determine the area densities, the objective is not to find estimators that predict the mass flow as precisely as possible, but to find area densities that resemble reality as close as they can.

If this is successful, material-specific mass flows can be determined, which enables the prediction of the product composition. Material-specific mass flows are calculated by multiplying the area flows with the determined area densities. A mathematical representation is depicted on the next page.

$$\begin{bmatrix} \dot{A}_{mat1,\ t1} * \rho_{mat1} & \dot{A}_{mat2,\ t1} * \rho_{mat2} & \dot{A}_{mat3,\ t1} * \rho_{mat3} & \cdots \\ \dot{A}_{mat1,\ t2} * \rho_{mat1} & \dot{A}_{mat2,\ t2} * \rho_{mat2} & \dot{A}_{mat3,\ t2} * \rho_{mat3} & \cdots \\ \dot{A}_{mat1,\ t3} * \rho_{mat1} & \dot{A}_{mat2,\ t3} * \rho_{mat2} & \dot{A}_{mat3,\ t3} * \rho_{mat3} & \cdots \\ \dot{A}_{mat1,\ t4} * \rho_{mat1} & \dot{A}_{mat2,\ t4} * \rho_{mat2} & \dot{A}_{mat3,\ t4} * \rho_{mat3} & \cdots \\ \dot{A}_{mat1,\ t5} * \rho_{mat1} & \dot{A}_{mat2,\ t5} * \rho_{mat2} & \dot{A}_{mat3,\ t5} * \rho_{mat3} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} = \begin{bmatrix} \dot{m}_{mat1,\ t1} & \dot{m}_{mat2,\ t1} & \dot{m}_{mat3,\ t1} & \cdots \\ \dot{m}_{mat1,\ t2} & \dot{m}_{mat2,\ t2} & \dot{m}_{mat3,\ t2} & \cdots \\ \dot{m}_{mat1,\ t3} & \dot{m}_{mat2,\ t3} & \dot{m}_{mat3,\ t3} & \cdots \\ \dot{m}_{mat1,\ t4} & \dot{m}_{mat2,\ t4} & \dot{m}_{mat3,\ t4} & \cdots \\ \dot{m}_{mat1,\ t5} & \dot{m}_{mat2,\ t5} & \dot{m}_{mat3,\ t5} & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix}$$

(5.2)

5.2.2 Data Pre-processing

To prepare the data for the application of the model, a histogram of NIR-scanner data and a scatter plot of NIR-scanner data together with belt weigher data was compiled. According to the belt weigher exploration, data below 0.3 [t/h] was excluded. Data points that were removed for the belt weighers were also excluded from the NIR-scanner data set. The respective plots are depicted in Figure 32.

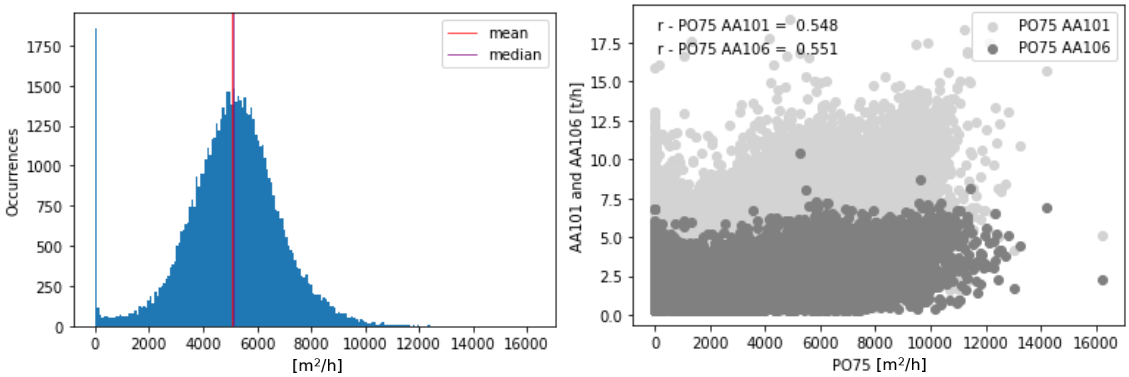

Figure 32: Total area flow histogram for PO75 and scatter plot of PO75 together with AA101 and AA106. NIR-scanner PO75 with all data points of belt weigher AA101 below 0.3 [t/h] removed.

Correlations extracted from the scatter plot are surprisingly low, with r values of 0.55 for PO75 and AA101, and 0.55 for PO75 and AA106. What furthermore can be seen, is that with the removal of empty data for AA101 and AA106, not all empty data for PO75 is removed. This is indicated by the left tail of the histogram as well as belt weigher data close to the y-axis of the scatter plot. To tackle this issue and to improve correlation, several thresholds to remove the left tail of PO75 were tried out. Good results were obtained for a threshold of 200 [m²/h]. Respective plots are depicted in Figure 33.

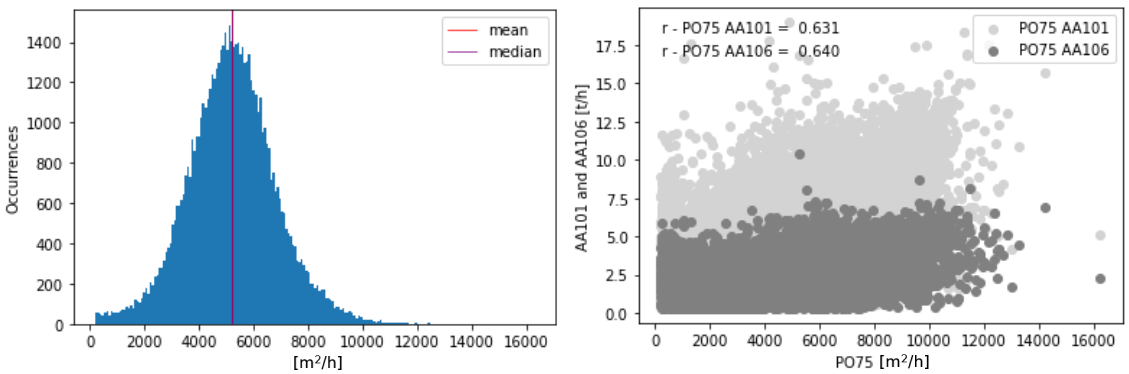

Figure 33: Total area flow histogram for PO75 and scatter plot of PO75 together with AA101 and AA106. NIR-scanner PO75 with all data points of belt weigher AA101 below 0.3 [t/h] removed and all PO75 area flow sums below 200 [m²/h] excluded.

It can be seen that the left tail of the histogram and a majority, of the data points close to the y-axis of the scatter plot is removed. Therefore, all empty data points of PO75 were excluded successfully. Nevertheless, correlations between PO75 and AA101 as well as PO75 and AA106 are still lower than expected. To further tackle this issue, the data was split into 200 bins by time and r was computed for each bin. This was done to check for temporal patterns in the data. The resulting plot can be found in Figure 34.

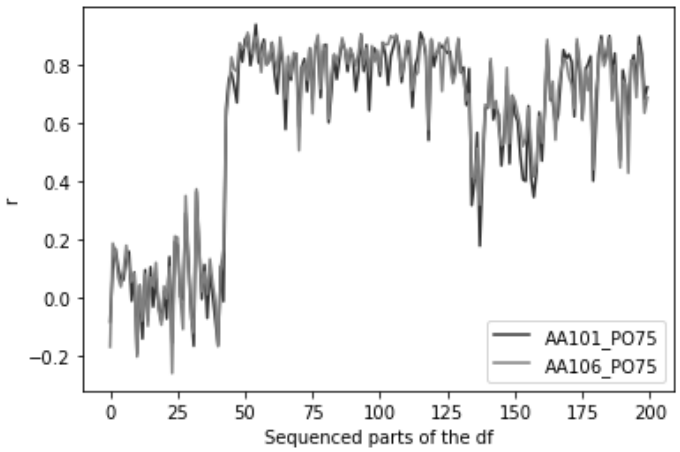

Figure 34: Pearson correlation coefficients for NIR-scanner PO75 with regards to belt weigher AA101 and AA106 with 200 bins compiled by time.

A clear temporal pattern is seen with a negative deviation up to bin number 45. Up to this bin, r oscillates around 0 with magnitudes of -0.2 and 0.4. Afterwards, the coefficient jumps to 0.8 and oscillates around this value. An exception of this is seen from bin 135 until bin 165 where the oscillation is shifted to 0.6. These patterns can be either explained by a malfunction of the belt weighers or the NIR-scanners. For the NIR-scanners, a change in the categorization program could be the reason, while for the belt weighers a changing measurement offset could be an explanation. To tackle this issue, it was decided to exclude all bins that have a correlation below 0.7. A histogram and scatter plot of the modified data is presented in Figure 35.

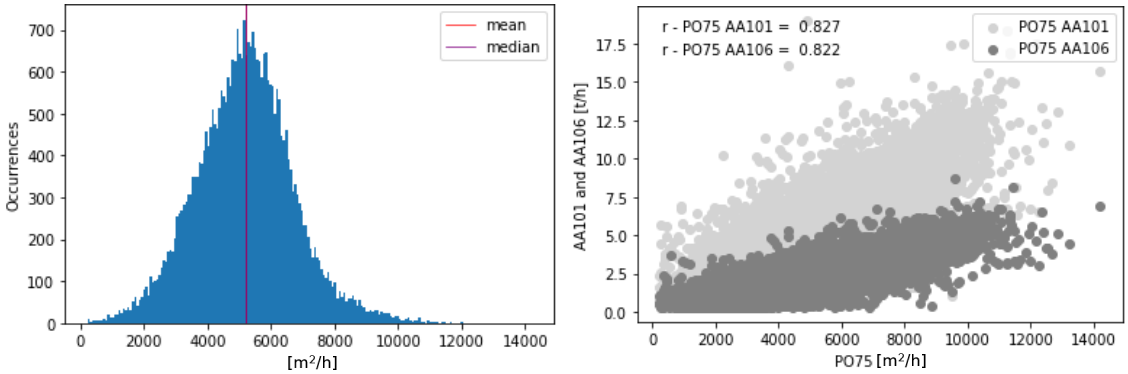

Figure 35: Total area flow histogram for PO75 and scatter plot of PO75 together with AA101 and AA106. NIR-scanner PO75 with all data points of belt weigher AA101 below 0.3 [t/h] removed, all PO75 area flow sums below 200 [m²/h] excluded and bins with correlations below 0.7 were not considered.

Although the amount of data points decreases dramatically from 65 thousand to 34 thousand, the correlation increases to expected r values of 0.83 for PO75 and AA101, as well as 0.82 for PO75 and AA106. Impossible values, where either the belt weigher or the scanners delivers no data, while the

other unit detects material, also decreases significantly. This becomes even more clear plotting a scatter plot with 0.5% coloration, where these instances for AA101 basically completely disappear.

In conclusion, the data pre-processing for OLS model application revealed a temporal correlation pattern for belt weigher data in combination with NIR-scanner area flows. Through this, an unexpected contribution to the answering of sub research question one is made. Complications for joint application of different data types were not expected. Nevertheless, difficulties could be overcome and new insights were gained.

5.2.3 OLS application

After proper pre-processing of the data, an 80/20 test training was realized. Afterwards, the OLS model was applied with NIR-scanner data from PO75 and belt weigher data from AA101. Results are depicted in Table 12.

Table 12: OLS results for area density prediction after exclusion of bins below r values of 0.7, belt weigher data below 0.3 [t/h] and total area flows on PO75 below 200 [m²/h].

	Regression Coefficient / Area density [kg/m²]	Standard Error [kg/m²]	95% interval
Constant	-245,296	24,148	[-292,628, -197,964]
BC	-0,153	0,166	[-0,479, 0,173]
BOARD_CT	18,930	1,461	[16,067, 21,794]
EPS	1,638	0,173	[1,298, 1,977]
OTHER_POLYMERS	8,83	0,669	[7,518, 10,142]
PAPER	5,874	0,261	[5,362, 6,386]
PET_BOTTLE	4,029	0,536	[2,979, 5,079]
PET_G	18,873	5,605	[7,888, 29,859]
PET_MONO_TRAY	0,36	0,595	[-0,807, 1,527]
PE_FILM	1,636	0,097	[1,446, 1,826]
PE_RIGID	-1,5	0,14	[-1,774, -1,226]
PP	1,699	0,171	[1,364, 2,033]
PP_FILM	-1,015	0,217	[-1,44, -0,59]
PS	-5,116	0,562	[-6,218, -4,014]
PVC	0,453	0,07	[0,316, 0,589]

At a first glance, materials like PVC and PE_FILM have area densities that lie in reasonable ranges and have small standard errors of 0.07 and 0.097. This would indicate expected deviation per m^2 of material of 70 grams and 97 grams, which could be acceptable. An especially negative example is represented by PET_MONO_TRAY. Here, the confidence interval ranges from -0.81 [kg/m^2] to 1.53 [kg/m^2]. This means that within the confidence interval it is unclear if the regression coefficient indicates a positive or a negative correlation. Apart from that, four negative regression coefficients occur. In this case, negative regression coefficients mean negative area densities. These are physically impossible, wherefore they should not occur.

A possible explanation for this could be multicollinearity, which is treated in the next subchapter. Furthermore, autocorrelation could influence the model result. Nevertheless, with a Durbin-Watson number of 2.07, there was no indication for further testing. Another explanation could be that one of the materials has a negative correlation with the belt weigher. Negative correlations are not expected, but

if they occur, they could be the reason for regression coefficients turning negative. To exclude this possibility, correlation of all materials from PO75, together with belt weigher AA101, are depicted in scatter plots and can be found in Figure 36.

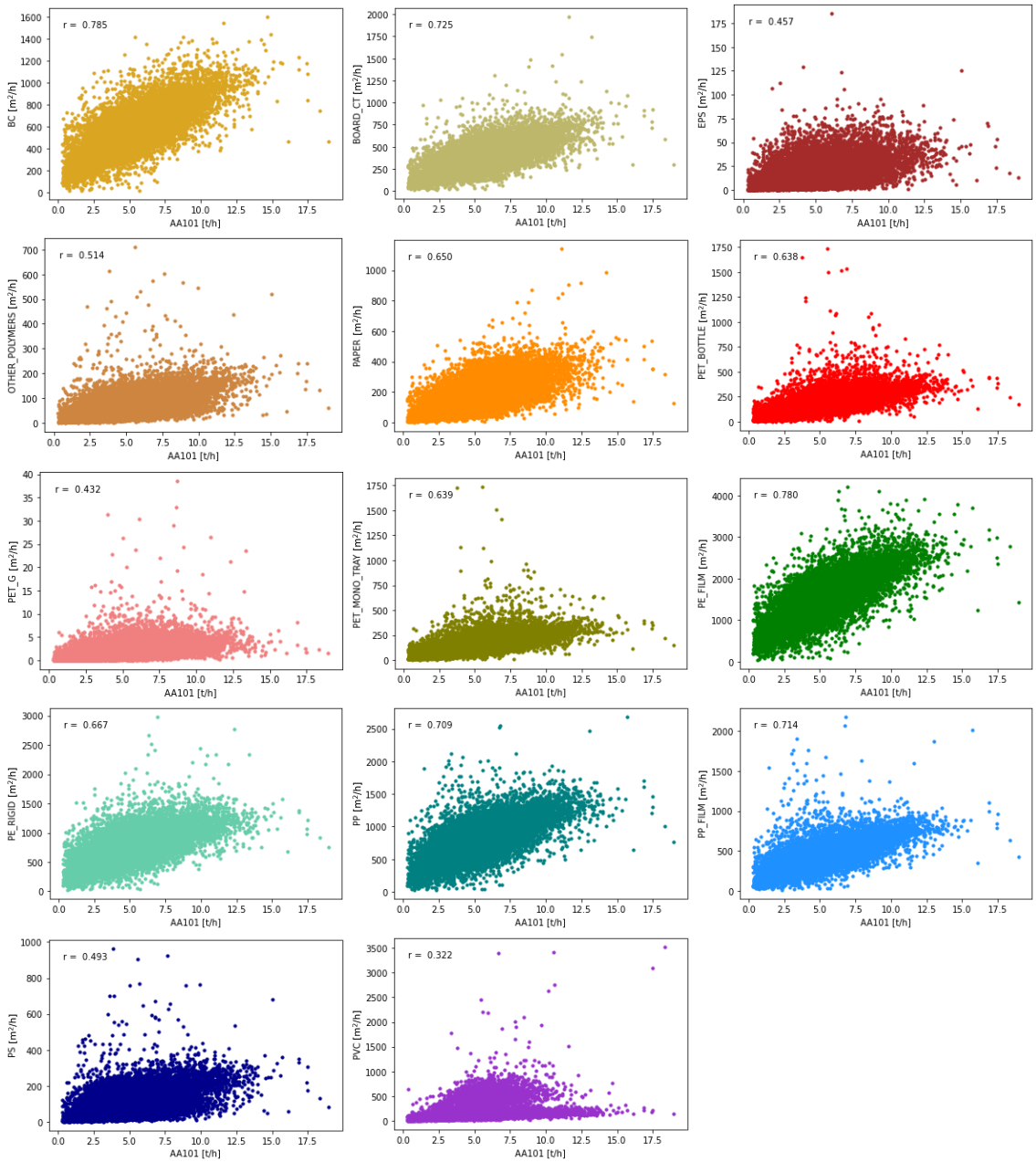

Figure 36: Scatter plots for materials from PO75 together with AA101 and an indication of the Pearson correlation coefficient on the top left corner.

Highest correlation is seen for BC and PE_FILM with 0.79 and 0.78. Lowest correlations are detected with 0.32 and 0.43 for PVC and PET_G. In total, all materials show positive correlation. Therefore, the hypothesis of a negative correlation between material and belt weigher leading to negative regression coefficients can be neglected.

Lastly, the OLS model could also be hampered by the violation of the normal distributed error assumption. To check for this a histogram of the errors is depicted in Figure 37.

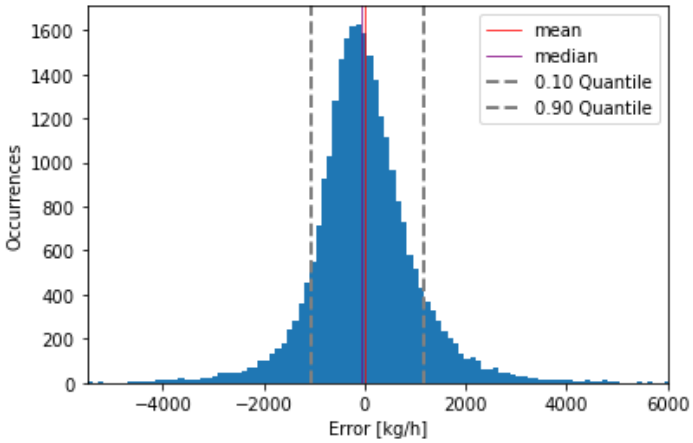

Figure 37: Error distribution of the initial OLS model with indications for the mean, median, 0.1 quantile and 0.9 quantile of the data with 200 bins.

Visibly, the assumptions of the normality of the errors is not violated. The errors are distributed around zero and the mean as well as the median align accordingly. The 0.1 and 0.9 quantile lie at -1072.1 [kg/h] and 1180.6 [kg/h]. Therefore, the error of the majority of the predicted datapoints lies in this area. With a mean of 5128 [m^2/h] for the belt weigher, this seems like a high range. The MAE for the model of 740 [kg/h] gives a slightly better perspective, but is still not excellent.

5.2.4 Multicollinearity

Multicollinearity can be detected through pairwise scatter plots or by compiling a correlation matrix for the independent variables. The second approach was chosen and the correlation matrix is depicted in Table 13.

Table 13: Correlation matrix for the independent variables of the OLS model without grouping. Dark colorations indicate higher values while light colorations indicate lower values. B_CT = BOARD_CT; OP = OTHER_POLYMERS; P = PAPER; PET_B = PET_BOTTLE; PET_M = PET_MONO_TRAY; PE_F = PE_FILM.

	B_CT	EPS	BC	OP	P	PET_B	PET_G	PET_M	PE_F	PE_R	PP	PP_F	PS	PVC
B_CT	1,00	0,52	0,82	0,52	0,89	0,48	0,30	0,51	0,67	0,52	0,59	0,67	0,54	0,24
EPS	0,52	1,00	0,39	0,76	0,33	0,43	0,30	0,44	0,38	0,36	0,45	0,45	0,88	-0,03
BC	0,82	0,39	1,00	0,45	0,79	0,53	0,36	0,53	0,90	0,80	0,70	0,71	0,43	0,44
OP	0,52	0,76	0,45	1,00	0,32	0,51	0,37	0,50	0,46	0,44	0,56	0,54	0,95	-0,05
P	0,89	0,33	0,79	0,32	1,00	0,32	0,19	0,35	0,55	0,36	0,36	0,46	0,36	0,37
PET_B	0,48	0,43	0,53	0,51	0,32	1,00	0,66	0,99	0,64	0,61	0,71	0,65	0,46	0,05
PET_G	0,30	0,30	0,36	0,37	0,19	0,66	1,00	0,64	0,45	0,45	0,49	0,45	0,34	0,01
PET_M	0,51	0,44	0,53	0,50	0,35	0,99	0,64	1,00	0,62	0,58	0,69	0,65	0,46	0,07
PE_F	0,67	0,38	0,90	0,46	0,55	0,64	0,45	0,62	1,00	0,93	0,89	0,85	0,42	0,41
PE_R	0,52	0,36	0,80	0,44	0,36	0,61	0,45	0,58	0,93	1,00	0,85	0,75	0,40	0,22
PP	0,59	0,45	0,70	0,56	0,36	0,71	0,49	0,69	0,89	0,85	1,00	0,96	0,51	0,17
PP_F	0,67	0,45	0,71	0,54	0,46	0,65	0,45	0,65	0,85	0,75	0,96	1,00	0,50	0,23
PS	0,54	0,88	0,43	0,95	0,36	0,46	0,34	0,46	0,42	0,40	0,51	0,50	1,00	-0,05
PVC	0,24	-0,03	0,44	-0,05	0,37	0,05	0,01	0,07	0,41	0,22	0,17	0,23	-0,05	1,00

What directly can be seen is that there is a high amount of correlation present between the independent variables. Looking at correlating categories, they follow an intuitive pattern, as they resemble the same material. As an example, PP has a correlation of 0.96 with PP_FILM and PE_Film has a correlation of 0.93 with PE_RIGID. The only categories that fall out of place are the high correlation of BC and PE_FILM with 0.9, and the low correlation of PET_G with PET_BOTTLE and PET_MONO_TRAY, with r values of 0.66 and 0.64, respectively. To get a better understanding of the correlation matrix, the mean correlation coefficient was calculated with 0.47. Furthermore, the occurrences of correlations above 0.85 were counted. With 10 occurrences they resemble 12% of the calculated correlations.

To treat the correlation between the categories, highly correlating materials were grouped sub sequentially. It was started by grouping PET_BOTTLE and PET_MONO_TRAY, as they had the highest correlation with 0.99. Next, PP and PP_FILM showed the correlation with the highest magnitude with a r value of 0.96. Accordingly, they were grouped. After this, still a very high correlation was seen for OTHER_POLYMERS and PS with 0.95. Additionally, EPS was added, as it has its two highest correlations with these materials and resembles a rather small area flow. Following this step PE and PE_FILM were grouped, with a correlation of 0.93. Although BC has a correlation of 0.9 with PE_FILM it was not added to avoid grouping of polyolefin and non-polyolefin material. As polyolefins are one of the materials that are tested for quality determination, BC cannot be added to the grouping. The reason for this is that it would become undistinguishable which part of the area density belongs to BC and which not. The highest correlation found at this point is between PAPER and BOARD_CT with 0.89. This time BC was added to the mix, as PAPER and BOARD_CT are the categories it has its best correlations with apart from PE. Lastly, PE and PP were aggregated to POLYOLEFINS as they still had a correlation of 0.88. For each step, the mean Pearson correlation coefficient, the count of correlations above 0.85 and the percentage of correlations above this threshold was compiled. Furthermore, the amount of negative regression coefficients and the number of categories is indicated. Results can be found in Table 14.

Table 14: Mean Pearson correlation coefficient, counts of correlations above 0.85 and threshold of these counts from the total number of calculated correlations for the OLS model data and its different grouping steps.

Grouping Steps	Mean r	Count > 0.85	Counts from total number [%]	Negative regression coefficients	Number of categories
Without grouping	0,47	10	12	4	14
PET grouped	0,46	9	13	4	13
PP grouped	0,44	6	10	3	12
PS and OTHER_POLYMERS grouped	0,45	4	10	2	10
PE grouped	0,42	3	9	1	8
CELLULOSICS grouped	0,38	1	6	0	7
POLYOLEFINS grouped	0,33	0	0	0	6

As expected, the mean Pearson correlation coefficient, the number of correlations above 0.85 and the percentage of these values decreases steadily for each grouping. One exception is seen for the PET grouping. Here, the number of categories and the number of instances with a correlation of greater 0.85 decreased by one. This leads to an increased share of values above 0.85 despite the negative trend. Nevertheless, after the last grouping no correlation above 0.85 can be found and regression coefficients

turn out positive with OLS modelling. In Table 15, the correlation matrix of the data after the last grouping is presented.

Table 15: Correlation matrix for the independent variables of the OLS after grouping. Dark colorations indicate higher values while light colorations indicate lower values.

	OTHER_POLYMERS	PET_G	PVC	PET	CELLULOSICS	PO
OTHER_POLYMERS	1,00	0,36	-0,05	0,49	0,49	0,49
PET_G	0,36	1,00	0,01	0,65	0,32	0,48
PVC	-0,05	0,01	1,00	0,06	0,38	0,30
PET	0,49	0,65	0,06	1,00	0,50	0,67
CELLULOSICS	0,49	0,32	0,38	0,50	1,00	0,74
PO	0,49	0,48	0,30	0,67	0,74	1,00

The highest correlation that is left is observed between CELLULOSICS and PO with 0.74. This is probably due to the high correlation between PE_FILM and BC, which were grouped into PO and CELLULOSICS. The next highest correlations are seen for PO and PET with 0.67, as well as with a r value of 0.65 for PET_G and PET. These correlations were considered acceptable, wherefore it was decided to continue the OLS model analysis with this setup. In Table 16, the results of the OLS modelling for the final grouping can be found.

Table 16: OLS results for area density prediction after exclusion of bins below r values of 0.7, belt weigher data below 0.3 [t/h], total area flows on PO75 below 200 [m²/h] and grouping.

	Regression Coefficient / Area density [kg/m²]	Standard Error [kg/m2]	95% interval
Constant	-395.958	23.703	[-442.416, -349.5]
OTHER_POLYMERS	1.527	0.074	[1.382, 1.673]
PET_G	18.882	5.619	[7.868, 29.895]
PVC	1.154	0.051	[1.055, 1.253]
PET	2.706	0.065	[2.579, 2.834]
CELLULOSICS	2.3	0.028	[2.241, 2.351]
PO	0.472	0.011	[0.45, 0.494]

The determined area densities have small standard errors between 0.011 [kg/m²] and 0.075 [kg/m²]. One exception for this is PET_G with a standard error of 5.619 [kg/m²]. While investigating where this deviation could result from, it was noticed that PET_G resembles only a minor share of the data. With a mean value of 1.9 [m²/h], compared to a mean of 5186 [m²/h] for the total area flows on PO75, PET_G only contributes 0.04% percent to the detected area. Therefore, it was decided to exclude it from the modelling. After PET_G was dropped from the data, the OLS modelling process was repeated. Results are depicted in Table 17.

Table 17: OLS results for area density prediction after exclusion of bins below r values of 0.7, belt weigher data below 0.3 [t/h], sum of areas on PO75 below 200 [m²/h], grouping and drop of PET_G data.

	Regression Coefficient / Area density [kg/m²]	Standard Error [kg/m²]	95% interval
Constant	-402.86	23.62	[-449.156, -356.572]
OTHER_POLYMERS	1.539	0.074	[1.394, 1.684]
PVC	1.1478	0.051	[1.049, 1.247]
PET	2.811	0.057	[2.699, 2.923]
CELLULOSICS	2.289	0.028	[2.234, 2.344]
PO	0.476	0.011	[0.454, 0.498]

After dropping PET_G, all standard errors of the material densities stay in the same range between 0.11 [kg/m²] and 0.074 [kg/m²]. For most of the categories, standard errors are unaltered, but for the constant, and for PET, they further decrease slightly to 23.62 [kg/m²] and 0.057 [kg/m²], from 23.7 [kg/m²] and 0.065 [kg/m²], respectively. In Figure 38, the error distribution for the newly trained OLS model can be found.

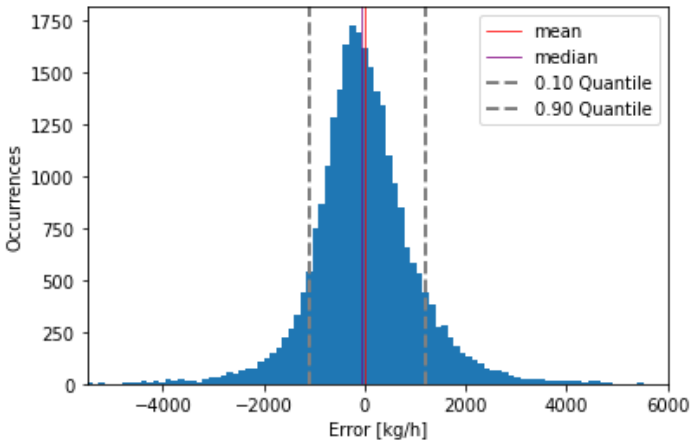

Figure 38: Error distribution of the OLS model with grouped data and indications for the mean, median, the 0.1 quantile and the 0.9 quantile of the data with 200 bins.

The error distribution is quite similar to the initial model, with a mean of zero and the 0.1 and 0.9 quantile at -1095.9 [kg/h] and 1206.3 [kg/h]. A slightly increased MAE of 757 [kg/h] can be observed, but still stays close to the original model. This can be explained by the nature of fixing multicollinearity. Through treated multicollinearity, a clear explanatory power is assigned to each regression coefficient, but the quality of the prediction is not necessarily increased.

5.2.5 Bootstrapping

To generate greater trust in the area densities, bootstrapping was applied. This was done to not only rely on the confidence interval provided by the initial sample. Bootstrapping is less prone to violation of assumptions due to its resampling nature. For the same reason, it is less prone to outliers. Simply said, a heavy influential outlier does not have to be drawn in each sample. Therefore, heavy influence of single data points would be revealed. The bootstrapping was done with 10,000 resampling rounds and results can be found in Figure 39.

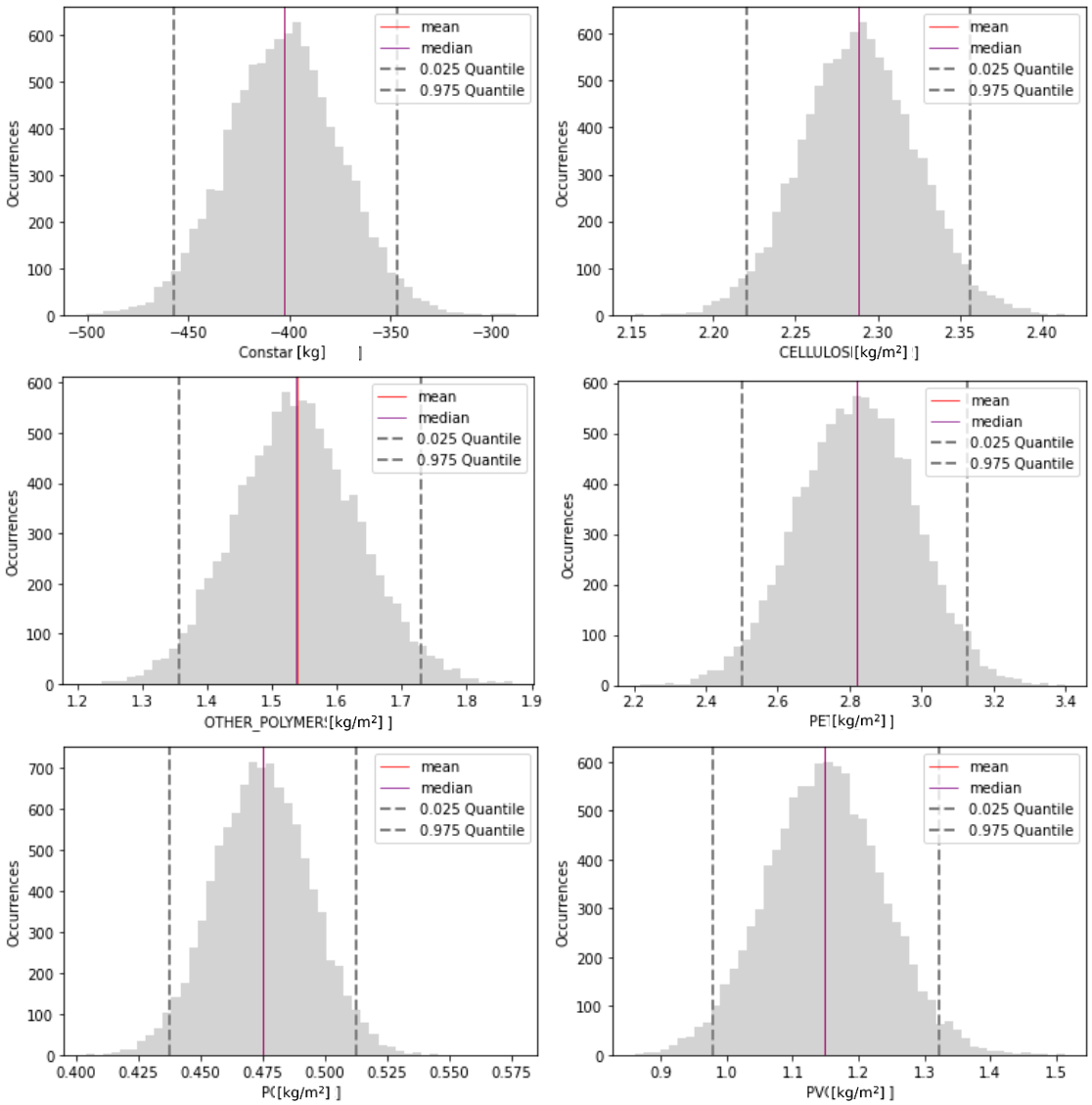

Figure 39: Bootstrapping results for the area densities and the constant of PO75 and AA101 after 10,000 resampling applications. The mean, the median and the 0.025 and 0.975 quantiles are indicated.

The resulting area density distributions for the bootstrapping align with the results from the OLS modelling and the determined confidence intervals. Greatest deviation from the mean of the distribution to the original regression coefficient can be found for the constant with a change from -402.86 [kg] to -402.73 [kg]. Second greatest change is found for PET with a deviation from 2.811 [kg/m^2] to 2.819 [kg/m^2]. The 0.025 and 0.975 quantiles can be seen as an equivalent to the 95% confidence interval. Compared to the confidence intervals, the quantiles of the bootstrapping are a little bit wider. Greatest deviation is observed for the constant where the quantile values are -456.94 [kg] and -347.01 [kg], while the confidence interval indicates -449.16 [kg] and -356.57 [kg]. For the area densities, PET and PVC show the greatest divergence, with quantile values of 2.5 [kg/m^2] and 3.13 [kg/m^2] for PET, and 0.98 [kg/m^2] and 1.32 [kg/m^2] for PVC, as well as confidence intervals of 2.7 [kg/m^2] and 2.92 [kg/m^2] for PET, and 1.05 [kg/m^2] and 1.25 [kg/m^2] for PVC. This indicates that through the bootstrapping additional variation in the data could be revealed, but apart from that no unexpected behaviour is found. The mean, median and quantiles for each material as well as the constant are depicted in Table 18.

Table 18: Summary of bootstrapping results presenting the mean, median, 0.025 and 0.975 quantile for the area densities determined with PO75 and AA101.

	Mean	Median	0.025 quantlile	0.975 quantile
Constant	-402.73	-402.64	-456.936	-347.01
CELLULOSICS	2.289	2.288	2.22	2.357
OTHER_POLYMERS	1.538	1.537	1.357	1.729
PET	2.819	2.821	2.501	3.127
PO	0.475	0.475	0.437	0.513
PVC	1.149	1.149	0.98	1.322

5.2.6 Testing and evaluation

To test the determined area densities, the test dataset was used. Area densities were multiplied with the area flows of the respective materials and plotted together with the belt weigher data. For better visibility, the data was aggregated in bins of 30 minutes. The plot can be found in Figure 40.

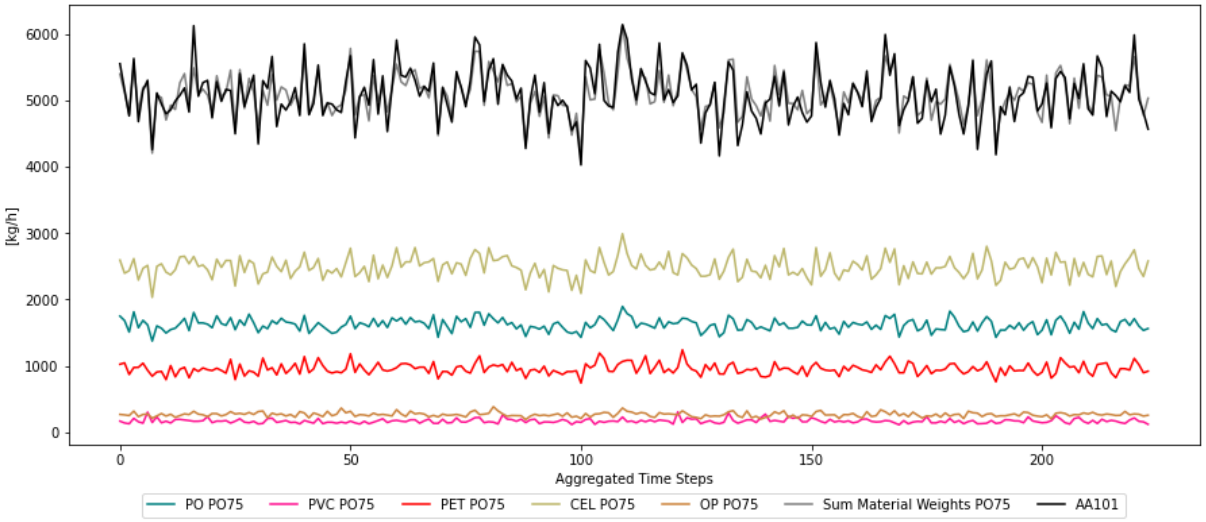

Figure 40: Comparison of actual belt weigher data from AA101 and mass flows compiled through area flows from PO75 with the help of area densities determined from PO75 and AA101, aggregation in bins of 30 minutes.

The fit appears accurate and follows the trends in the data. Most of the peaks are matched by the prediction with exemplary exceptions around time step 20, 35 and 215. In these occasions, peaks were predicted but the data showed a rather small or no peak. Through the aggregation of the data, the MAE decreases from 760.7 [kg/h] to 141 [kg/h]. This resembles a mean relative error (MRE) of 16% and 3%. This happens as due to the aggregation positive and negative deviations balance each other out. Therefore, the prediction appears more accurate in greater time step aggregations. To gain a better understanding of the performance of the model, error distributions for the aggregated and unaggregated data is presented in Figure 41.

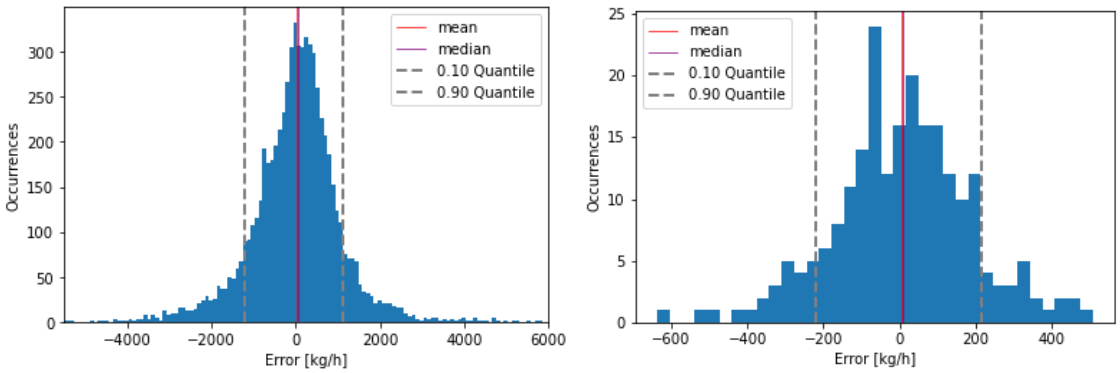

Figure 41: Error distribution of the OLS model with no aggregation (left) and aggregation of 30-minute periods (right) for test data application together with indications for the mean, median, the 0.1 quantile and the 0.9 quantile of the data.

The 0.1 quantile is measured at -1201.8 [kg/h] while the 0.9 quantile lies at 1105.4 [kg/h]. Compared to the quantiles of the trainings data, a shift can be observed towards greater values for the negative side of the distribution, while for the training data the 0.9 quantiles tended to be of greater magnitude. Apart from that, the errors are normally distributed. The shape of the errors of the aggregated data does not follow a normal distribution as clearly as the unaggregated data does, but its normality is still visible. This is due to the decreased amount of data points, wherefore outliers gain greater influence. The 0.1 and the 0.9 quantile are determined with -218.7 [kg/h] and 214.8 [kg/h].

Gathering this data, a clear picture of the abilities of the model can be drawn. Results from the aggregation show that with higher aggregation higher accuracy can be obtained. Therefore, the question for the need of granularity for the final model application is raised. Apart from that, in the final quality prediction, it has to be shown if the determined accuracy is precise enough for a reliable quality prediction. Furthermore, the obtained constant symbolizes the part of the model that could not be assigned to a specific material, wherefore a strategy to treat this part of the data needs to be developed.

5.2.7 Generalizability

In a next step, the generalizability of the area density was tested. For this, area densities determined with PO75 and AA101 were applied to belt weigher and NIR-scanner data of AA106 and PO75C. Results are presented in Figure 42.

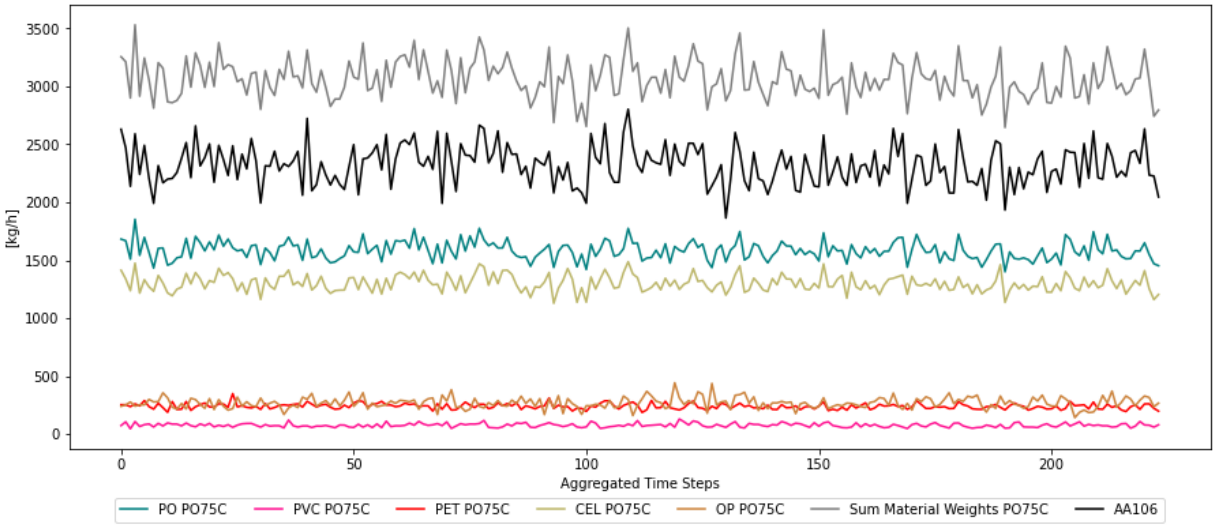

Figure 42: Comparison of actual belt weigher AA106 data and data compiled through area flows from PO75C and determined area densities from PO75 in combination with AA101.

It can be seen that, by applying the area densities determined with PO75 and AA101 to predict AA106, values are heavily overestimated. The MAE for this case is indicated with 743.9 [kg/h], for the aggregated data, and with 862 [kg/h], for the unaggregated data. Furthermore, it can be deducted from the plot that a strictly positive deviation is present. This indicates a positive offset. Although it is inconvenient that the determined area densities are not applicable across multiple NIR-scanners, it is an expected finding.

Due to the separation of targeted material, the area densities of the material flow will change. This is due to the fact that not 100% of the targeted material will be transferred to the following scanner. The particles that reach the next scanner are biased through the separation abilities of the previous scanner. This means that the first scanner will separate particles with certain properties better than others. Therefore, particles with a property set influenced by the separation itself will reach the next scanner. In return, the particles on the next scanner have a different set of properties than on the previous one.

5.2.8 Fitting of PO75C and AA106

Due to the prohibited generalizability of area densities from PO75 and AA101, the area densities from PO75C and AA106 are fitted and analysed in this subchapter. As the principles of this fitting procedure were explained in depth in the previous subchapters, an abbreviated version is presented in this section. Accordingly, a similar grouping procedure as for PO75 was applied and the results of the OLS modelling for the final grouping can be found in Table 19.

Table 19: OLS results for area density prediction after exclusion of bins below r values of 0.7, belt weigher data below 0.3 [t/h], sum of areas on PO75C below 200 [m²/h], grouping and drop of PET_G data.

	Regression Coefficient / Area density [kg/m²]	Standard Error [kg/m²]	95% interval
Constant	-298.08	15.539	[-328.537, -267.623]
OTHER_POLYMERS	0.33	0.022	[0.287, 0.373]
PVC	0.25	0.053	[0.146, 0.354]
PET	3.814	0.105	[3.608, 4.02]
CELLULOSICS	0.811	0.044	[0.724, 0.897]
PO	0.534	0.009	[0.517, 0.551]

Standard errors lie in a range from 0.009 [kg/m2] to 0.105 [kg/m^2], for the materials, and a standard error of 15.54 [kg/m^2] is obtained for the constant. For OTHER_POLYMERS and PO, the standard error outperforms the previous fit of the area densities. Regarding PVC, PET and CELLULOSICS, the other area densities have smaller standard errors. Nevertheless, in between the two fits standard errors are small and confidence intervals are acceptably narrow. To gain further understanding of the behaviour of the determined area densities, bootstrapping was applied. Here 10,000 resampling rounds were conducted and results are presented in Figure 43.

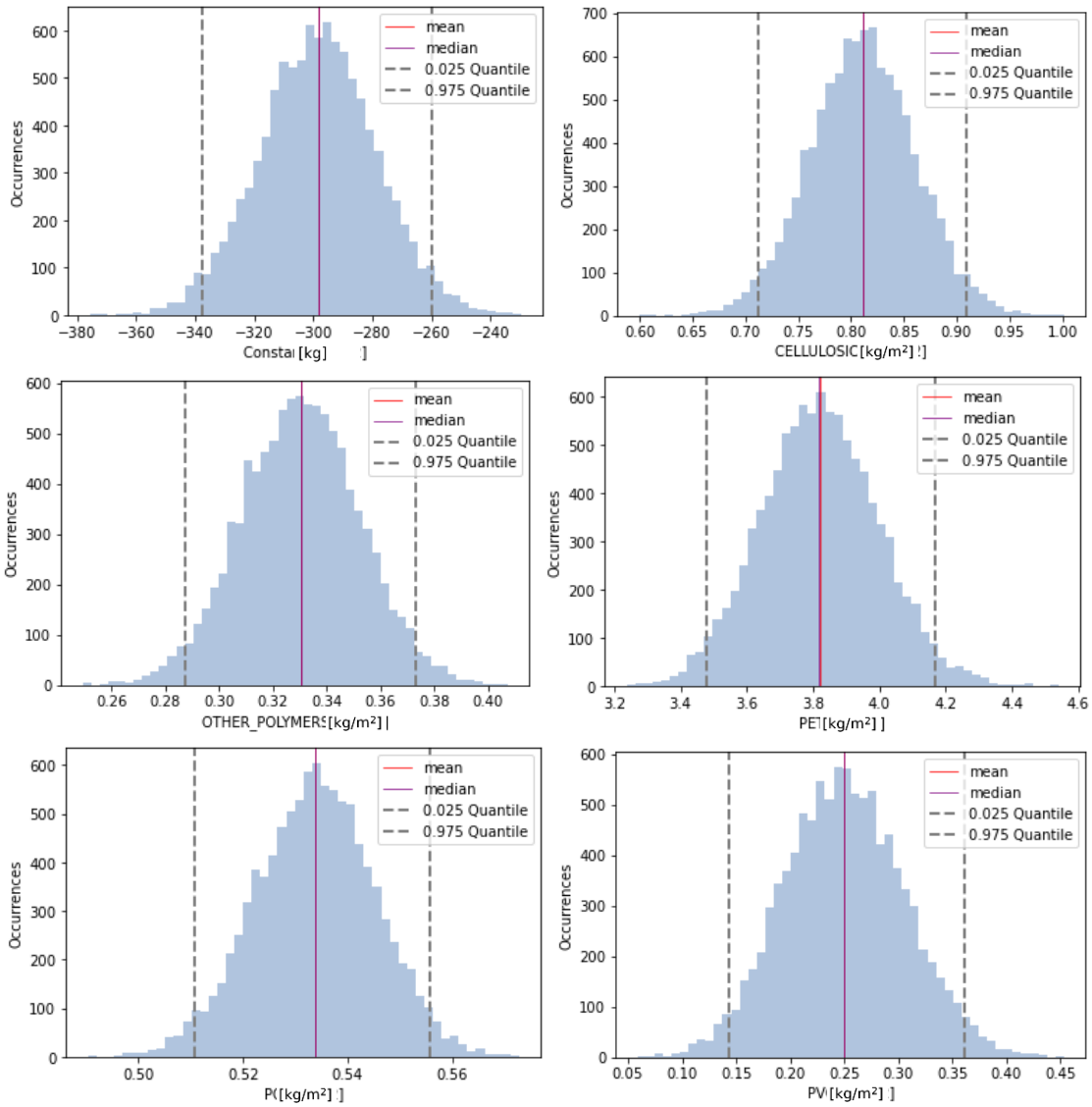

Figure 43: Bootstrapping results for the area densities and the constant of PO75 and AA101 after 10,000 resampling applications, the mean, the median and the 0.025 and 0.975 quantiles are indicated.

Also, for PO75C and AA106, the results align with the determined area densities from the OLS model and their confidence intervals. Especially, the mean and median overlap accurately with the area densities from the OLS model. For better overview, all quantiles, means and medians from the bootstrapping are presented in Table 20.

Table 20: Summary of bootstrapping results presenting the mean, median, 0.025 and 0.975 quantile for the area densities determined with PO75C and AA106.

	Mean	Median	0.025 quantlile	0.975 quantile
Constant	-298.41	-298.34	-337.59	-259.67
CELLULOSICS	0.811	0.811	0.712	0.909
OTHER_POLYMERS	0.330	0.331	0.287	0.373
PET	3.819	3.819	3.481	4.166
PO	0.534	0.534	0.511	0.556
PVC	0.25	0.249	0.143	0.362

Compared to the confidence intervals, the quantiles are wider. An exception is resembled by OTHER_POLYMERS, which has exactly the same value as the confidence intervals. Apart from that, the changes are small with below 0.02 [kg/m^2], for the materials, and below 10 [kg] for the constant. Delving into the evaluation, compiled mass flows were plotted together with measured belt weigher data against time. To do this, area flows have been multiplied with the respective area densities and summed up for comparison with the belt weigher. For better visibility, the data was aggregated into bins of 30 minutes. The plot is depicted in Figure 44.

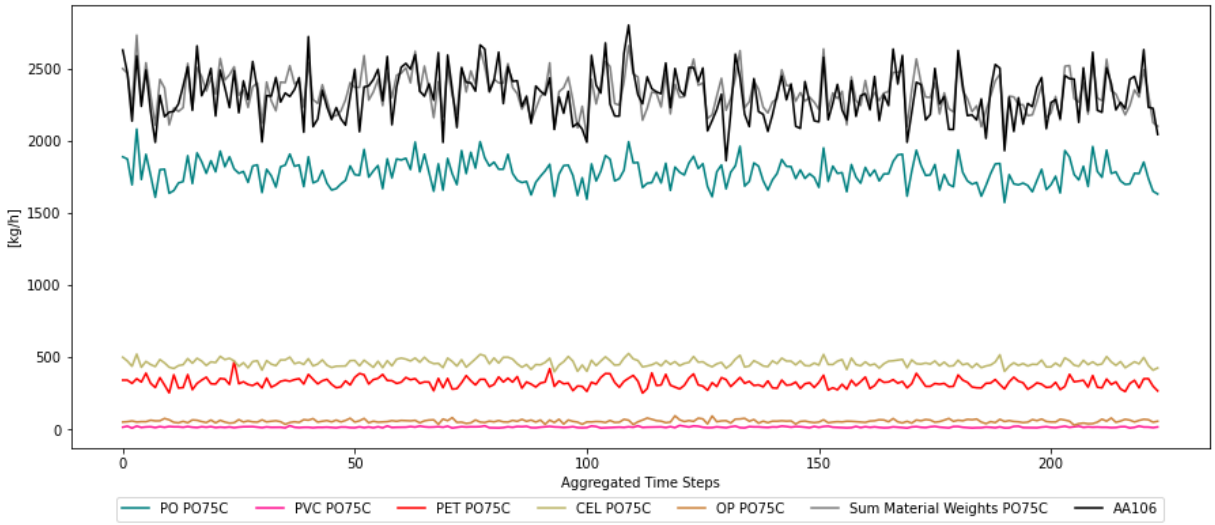

Figure 44: Comparison of actual belt weigher data from AA106 and mass flows compiled through area flows from PO75C with the help of area densities determined from PO75C and AA106.

By visual inspection, a good fit of the data can be observed. In some cases, the prediction deviates like for bin 40 and 55, but overall the majority of positive and negative peaks is matched by the model. Following up with a more parameter driven evaluation, the MAE is 92.6 [kg/h] for the aggregated data and 486.7 [kg/h] for the non-aggregated case. Lastly, the error distributions were analysed and are presented in Figure 45.

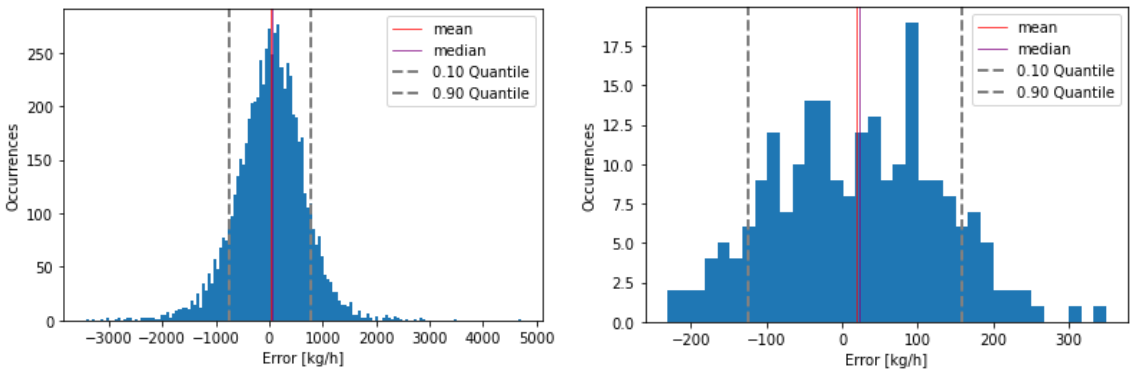

Figure 45: Error distribution of the OLS model for PO75C with no aggregation (left, 150 bins) and aggregation of 30-minute periods (right, 35 bins) for unseen data, indications for the mean, median, the 0.1 quantile and the 0.9 quantile of the data.

Interestingly, the mean of the error distribution is shifted for both versions by 19.6 into the positive direction of the distribution. This is most probably due to the application of the model to the test data, as for the trainings data the mean is centred around 0 [kg/h]. The quantiles for the non-aggregated results lie at -758.3 [kg/h] and 765.4 [kg/h], while for the aggregated version they are determined with -124.7 [kg/h] and 158.8 [kg/h].

Concluding, the modelling of area densities using OLS was demonstrated and evaluated. This contributes to answering the second sub research question, which focuses on modelling key components necessary for quality prediction of the agglomeration line product. Furthermore, insights were gained about the generalizability of the area density, which inconveniently is not given.

5.3 Machine Learning – Area Flow prediction

To compile another puzzle piece for the quality prediction of the agglomeration product, predicting the area flows on PO75C based on the area flows from PO75C is of interest. For this, a ML approach was chosen, as non-linear behaviour for the prediction is expected. Furthermore, through this, belt occupation data can be easily included without derivation of physical correlations between belt occupation and material transference. Lastly, the results of this subchapter will contribute to answering sub research question two.

5.3.1 Data Pre-processing

Inspired by the data exploration different pre-processing options were applied and are elucidated in Table 21.

Table 21: Pre-processing versions applied to prepare the data for machine learning model training and application.

Version	Description
V1	NaN dropped
V2	NaN and zeros dropped
V3	NaN and zeros dropped, area sum thresholds (PO75: 1000/12500; PO75C: 1250/10000) applied
V4	NaN and zeros dropped, two additional materials from PO75 left in
V5	NaN and zeros dropped, two additional materials from PO75 left in, area sum thresholds
V6	NaN and zeros dropped, two additional materials from PO75 left in, AA101 and AA106 data included
V7	NaN and zeros dropped, two additional materials from PO75 left in, AA101 and AA106 data included, area sum thresholds
V8	NaN and zeros dropped, two additional materials from PO75 left in, AA101 and AA106 data included, belt occupation indicator included
V9	NaN and zeros dropped, two additional materials from PO75 left in, AA101 and AA106 data included, belt occupation indicator included, area sum thresholds

As can be seen in the Table, the versions are gradually constructed. The derivation of the total area flow thresholds is explained in subchapter 5.1.2. All sums that lie below the first value or above the second value are excluded for all data. This means that the entire data point is removed, to avoid NaN handling and other numerical problems. With two additional materials, two further material categories that are included for PO75 but are not present on PO75C are meant. These categories are PET_G and BOARD_CT. Initially, they were left out for consistency in between the scanners. With the belt occupation indicator, the occupation counts explained in subchapter 5.1.3 are referred to. Finally, with inclusion of AA101 and AA106, the incorporation of the available belt weigher data is meant.

Apart from that, after the pre-processing of the data, an 80/20 train test split is applied.

5.3.2 Model Try Out and Selection

For the broad model try out, the following models were tested: Decision Tree (DT), Extra Tree (ET), Random Forest (RF), Gradient Boosting (GB), K-Nearest Neighbours (KNN), Bagging Regressor (BR), Ridge Regression (RR), Elastic-Net Regression (ENR) and Multi-Layer Perceptron (MLP). Model performance is evaluated based on the MAE, the MSE and the R^2 value. Indicator calculation is done for the results of the model that was trained on the training data and afterwards applied to the test data. The three indicators were plotted for all models and all versions with one plot per indicator. Materials are indicated separately to develop insight into the contributions of each material to the indicator. Results can be observed in Figure 46, 48 and 49.

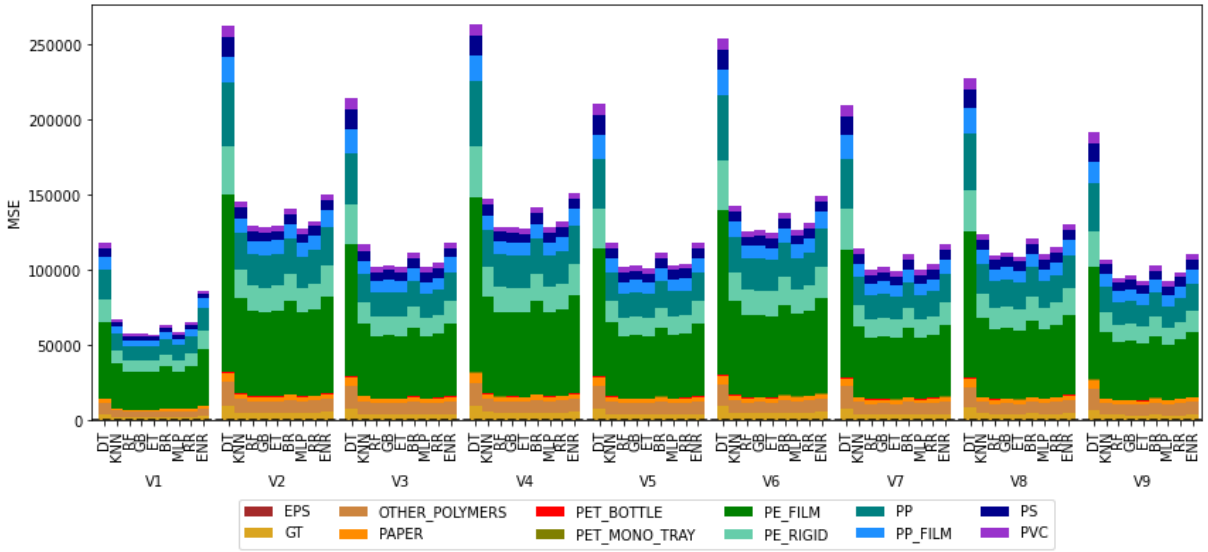

Figure 46: MSE for all models participating in the try out, grouping by data pre-processing version and contribution of each material presented as a stacked bar. DT = Decision Tree; ET = Extra Tree; RF = Random Forest; GB = Gradient Boosting; KNN = K-Nearest Neighbours; BR = Bagging Regressor; RR = Ridge Regression; ENR = Elastic-Net Regression; MLP = Multi-Layer Perceptron.

For the MSE, smaller values indicate a better fit of the model. As the errors are squared, emphasis is laid on large errors. Therefore, the MSE is also a good indication for a constantly well performing model. What directly can be seen, is that V1 is performing especially well compared to the other versions. This is due to the large number of zeros and values close to zero in the data. Therefore, the prediction is not necessarily better but errors are smaller. As shown in the data exploration, data points with this small area flow occurrence introduce great randomness. Therefore, together with the described distortion of the model evaluation, it was decided to remove V1 from the selection process. An updated plot can be found in Figure 47.

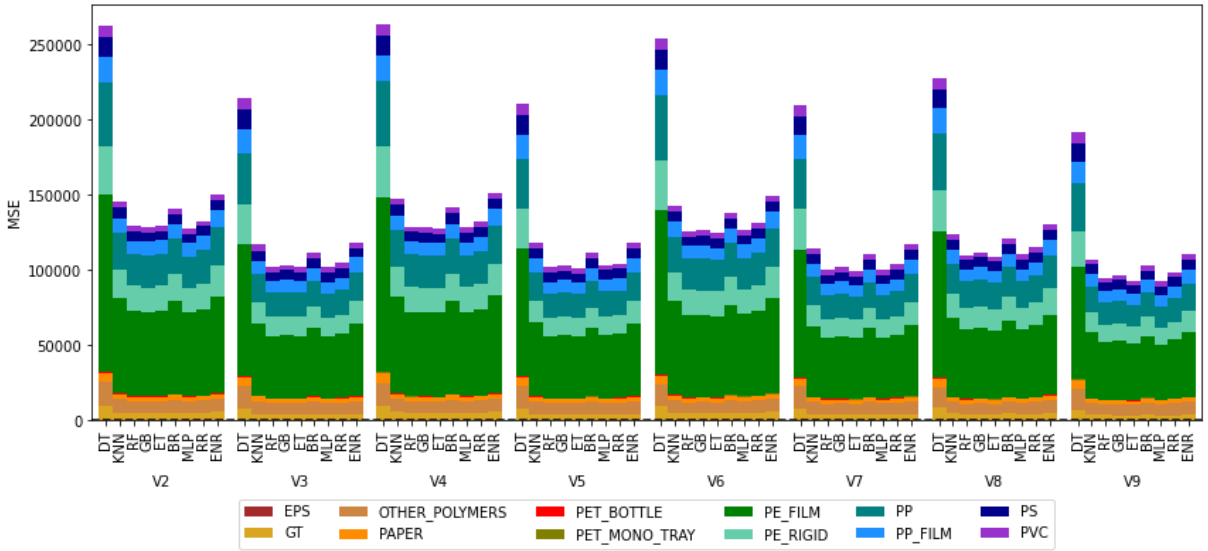

Figure 47: MSE for all models participating in the try out, grouping by data pre-processing version with V1 excluded and contribution of each material presented as a stacked bar. DT = Decision Tree; ET = Extra Tree; RF = Random Forest; GB = Gradient Boosting; KNN = K-Nearest Neighbours; BR = Bagging Regressor; RR = Ridge Regression; ENR = Elastic-Net Regression; MLP = Multi-Layer Perceptron.

In the updated plot, it can be observed that performance oscillates between odd and even version numbers. Furthermore, an improving trend for later version is seen. For MSE, odd version numbers perform better. This are the versions that have the thresholds for the total area flows included. Interpreting this means that these versions either result in more outlier resistant models or have less outliers in their data initially. Apart from that, the best performing model (MLP) decreases from an MSE of 127,89 [m^4/h^2] in V2 to 92,58 [m^4/h^2] in V9. Furthermore, first trends of the model performance can be derived. In all version, the DT model performs worst, while RF, GB, ET, MLP and RR show the best performances in a tight range. Apart from that, biggest contributions to the MSE come from PE_FILM, PE_RIGID and PP. The two smallest contribution are seen by PET_BOTTLE and EPS. This aligns with the general occurrence of the material, as PE_FILM, PE_RIGID and PP have the highest mean magnitude of area flows, while PET_BOTTLE and EPS are on the lower end. To investigate model performance, also on a less outlier prone basis, the MAE was computed next. Results are depicted in Figure 48.

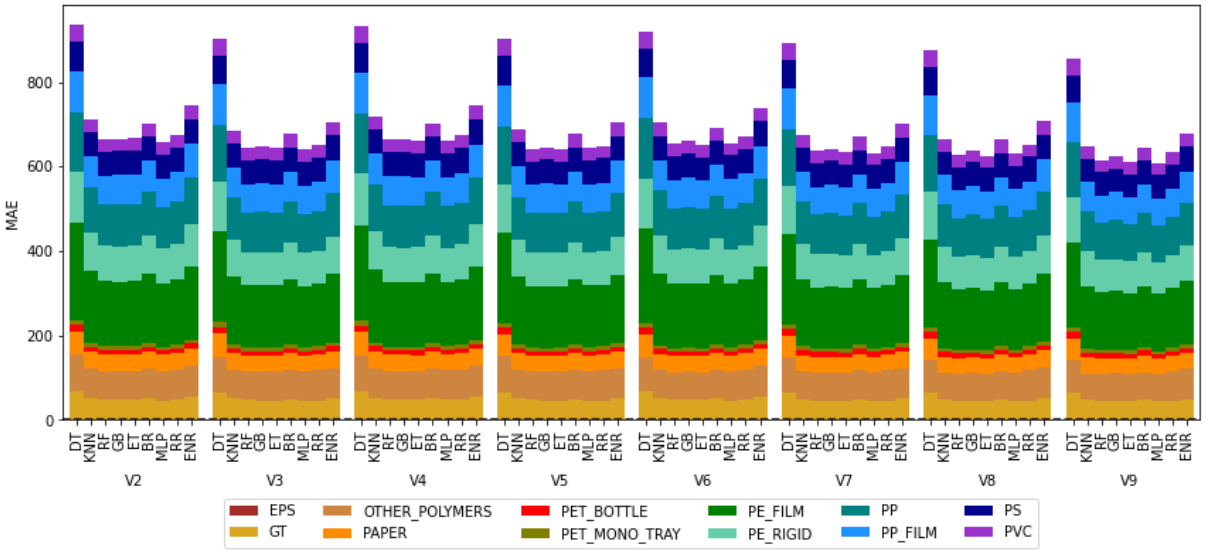

Figure 48: MAE for all models participating in the try out, grouping by data pre-processing version with V1 excluded and contribution of each material presented as a stacked bar. DT = Decision Tree; ET = Extra Tree; RF = Random Forest; GB = Gradient Boosting; KNN = K-Nearest Neighbours; BR = Bagging Regressor; RR = Ridge Regression; ENR = Elastic-Net Regression; MLP = Multi-Layer Perceptron.

A similar but less pronounced trend between the versions can be observed. The odd versions perform again better. These are the versions, which include the thresholds for the summed area flows. Furthermore, this means that these versions show a very stable performance. Together with the good performance for MSE, this indicates that the most robust models will result out of the odd versions data pre-processing. Apart from that, the greatest contributions to the MAE stem again from PE_FILM, PE_RIGID and PP. Nevertheless, the contribution of PP_FILM, OTHER_POLYMERS and PS increased compared to the MSE. This is due to the nature of the squaring during the compilation of the MSE and indicates that these categories have more consistent errors with less extremes. The best performing models, over all versions, are again RF, GB, ET, MLP and RR. Of these models MLP outperforms the rest and is decreasing from 657.1 [m^2/h] to 608.4 [m^2/h] from V2 to V9.

Moving away from analysing the magnitude of the errors and the stability of the model, R^2 was compiled for all versions and materials. The regarding plot can be found in Figure 49.

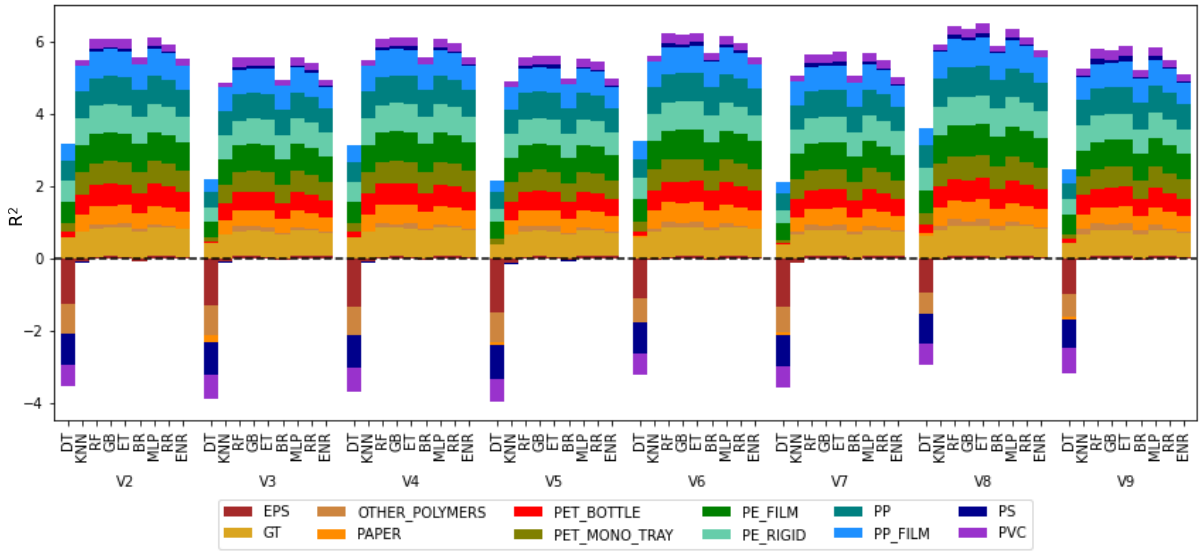

Figure 49: R^2 for all models participating in the try out, grouping by data pre-processing version with V1 excluded and contribution of each material presented as a stacked bar. DT = Decision Tree; ET = Extra Tree; RF = Random Forest; GB = Gradient Boosting; KNN = K-Nearest Neighbours; BR = Bagging Regressor; RR = Ridge Regression; ENR = Elastic-Net Regression; MLP = Multi-Layer Perceptron.

For R^2, higher values indicate a better performance. An R^2 value of 1 means complete explanation of the variability in the data and the capturing of all patterns that are present. An oscillation of performance between odd and even versions can be seen. Furthermore, better performance for later versions is observed. Looking back at the MSE and MAE, it seems contradictory that this time the even versions perform better. This means that versions that perform better for R^2 perform worse for MSE, as well as MAE, and the other way around. A possible explanation of that is that the even versions manage to better detect patterns in the data, but at the same time produce larger errors. A simple example for this would be the comparison of an offset, with regards to the measured values, and an oscillation around them. If the oscillation is small enough, it would produce a smaller error but would not capture the present pattern as well as the offset would do. Apart from that, RF, GB, ET, MLP and RR are again the best performing models and out of this group ET displays the largest R^2 value. Comparing the R^2 value of ET between the versions it increases from 5.58 to 6.5 from V3 to V8.

Delving into the model selection, a governing indicator has to be determined. The predicted area by the ML model will be used together with the modelled area densities to determine mass flows for product composition and quality prediction. Therefore, smaller errors are valued more over better detected patterns. As a result, the MAE and the MSE will be preferred over R^2. Using the same argument, it will be focused on the MAE instead of the MSE, as smaller general errors are preferred over robustness to outliers. Accordingly, V7 and V9 were picked for further analysis.

To bring material-specific behaviour better into play, material-specific model performance was plotted for V7 and V9. In Figure 50, the regarding plots for PE_FILM, PE_RIGID, PP, PP_FILM and PVC are presented. This list of materials was picked, as the quality determination of the agglomeration product is guided by these materials. Plots for all materials can be found in Appendix 7.

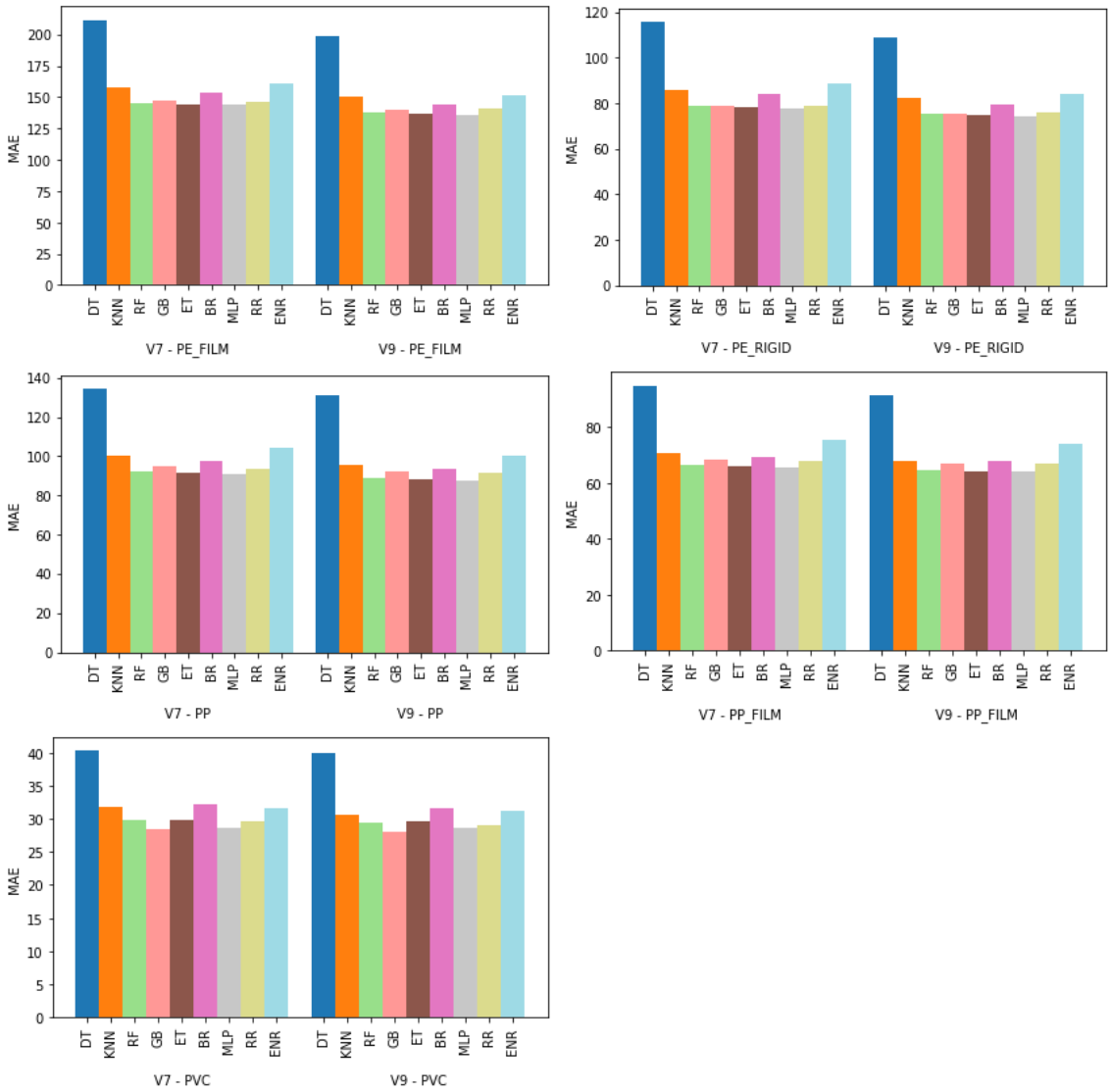

Figure 50: Model performance measured in MAE for V7 and V9 pre-processing and relevant materials regarding the quality prediction of the agglomeration product.

As seen in the previous analyses, there are five models that perform best across the scrutinized materials. These models are RF, GB, ET, MLP and RR. It can be seen that ET and MLP perform best, with an exception for PVC. Here, GB has a better MAE. For the remainder of the cases, GB together with RR perform worst within the top group. Apart from that, the DT model performs worst for all materials.

The DT model fits a single decision tree on the data. Through the bad performance, it is shown that the modelling task is too complex for this approach. An explanation could be that the data is too widespread for a single decision tree. Therefore, no split route can be found that captures the variability of the data. Furthermore, the approach could be hampered by the multicollinearity present in the data. Here, the model will spread splitting decisions randomly across correlated features. This happens, as these features share explanatory power, which increases uncertainty. From the best performing models RF, GB and ET represent decision tree-based ensemble models. This speaks for the fact that the use of several decision trees in combination can keep up with the patterns present in the data. An exemption

to the strong performance of decision tree-based ensemble models is the BR model. The BR model is also a decision tree-based ensemble model, but does not show the performance of RF, GB and ET. This is due to the fact that BR does not introduce additional randomness to the model, like RF and ET. Furthermore, it does not improve the model by fitting the new decision trees on the result of the loss function, like GB does. Instead, it only introduces variation through bootstrapping, which does not seem to be enough to capture the data in the patterns. Comparing BG to the performance of the DT model, it is shown that the bootstrapping approach can balance variability and potentially counteracts multicollinearity. Nevertheless, it is outperformed by ensemble models that incorporate additional randomness or optimize splits based on previous errors.

The RR model is part of the best performing group and represents a linear regression approach. This speaks for linear trends that are present in the data and fits previous observations. Models based on singular decision trees struggle with this type of relationship. This happens, as they split the data at different points and therewith create piecewise constant predictions instead of linear patterns. As also ENR represents a linear regression model, they question of why RR performs better than ENR is risen. Compared to RR, ENR adds another penalty term to the loss function. The additional penalty term is the sum of all absolute weights. This effectively performs feature selection, as the model attempts to set some feature weights to zero in order to minimize the sum of the absolute weights. At first, the better performance of RR shows that all features are relevant for the workings of the model. This is proven by the fact, that ENR inherently tends to exclude features and performs worse than RR. Furthermore, due to multicollinearity, ENR is likely to eliminate all but one feature from a group of correlated features. As high multicollinearity was shown in subchapter 5.2.4, the better performance of RR can be explained.

To choose models for the hyperparameter tuning, additionally to the described analysis, the presence of different model architecture was considered. Therefore, MLP and RR are set for the hyperparameter tuning. This was done, as they represent neural network and linear regression architectures. GB, ET and RF are all tree-based ensemble model. Although GB performs worst for four out of five materials in the ensemble model group, it is selected for hyperparameter tuning. This was done, as the greatest improvement through tuning is expected for this model.

Lastly, the XGBoost model was added to the hyperparameter tuning. This was done, as it resembles an improved version of the GB model and was not considered upfront by accident.

5.3.3 Hyperparameter tuning and cross validation

The used hyperparameter spaces for the hyperparameter tuning can be found in Table 22. Explanation of the functioning for each hyperparameter and the reasoning for the selected search space are explained in subchapter 2.6.3 and 4.6.2. The hyperparameter tuning was conducted via grid search.

Table 22: Hyperparameter space for grid search for Ridge Regression, MLP, Gradient Boosting and XGBoost.

Model	Hyperparameter	Values applied in grid search
Ridge Regression	alpha	[0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 500, 1000, 5000, 10000]
	solver	[svd, cholesky, lsqr, sag, sparse_cg]
MLP	hidden_layer_size	1 layer: [50, 100, 150] 2 layers: [[50, 50], [100, 100], [150, 150]] 3 layers: [[50, 50, 50], [100, 100, 100], [150, 150, 150]]
	alpha	[0.001, 0.01, 0.1]
	activation	[tanh, relu, logistic]
Gradient Boosting	n_estimators	[50, 250, 500]
	learning_rate	[0.05, 0.15, 0.3]
	max_feature	[7, 14, 21]
	max_depth	[2, 10, 18]
	min_samples_split	[2, 50, 100]
XGBoost	n_estimators	[50, 175, 399]
	learning_rate	[0.1, 1, 2]
	max_depth	[1, 8, 15]
	gamma	[0, 0.25, 0.5]
	lambda	[0, 0.15, 0.3]

Together with the hyperparameter tuning cross validation was implemented. To keep computational cost at bay, a 5-fold cross validation was applied.

For Gradient Boosting, best results were achieved with a learning rate of 0.05, a max_depth of 2, a maximum of 7 features, a minimum sample split of 2, and 50 estimators. As a result, a MAE of 56.9 [m2/h], an MSE of 9446.7 [m^4/h^2] and a R^2 of 0.422 were obtained. Interestingly enough, running the Gradient Boosting algorithm with its default settings leads to better results with 51.9 [m^2/h] for the MAE, 8076.4 [m^4/h^2] for the MSE and 0.48 for R^2. This showcases the weaknesses of grid search, as only a fixed set of combinations is searched. Nevertheless, the information is obtained that an ideal hyperparameter setting can be found around the default values. This is due to the fact that the indicators show better performance for the default values. Therefore, a local or global optimum should be present around these values. This optimum could be identified by a more specific grid search around these values or an alternative but more advanced hyperparameter tuning methodology.

Regarding the MLP model, best results were obtained with the rectified linear unit function as activation function, an alpha value of 0.1, and three hidden layers with 100 neurons each. This results into a MAE of 51.4 [m^2/h], an MSE of 8191.3 [m^4/h^2] and a R^2 of 0.435. Also, for the MLP model, the hyperparameter tuned version is outperformed by the default version of the model. The default version has a MAE of 51.1 [m^2/h], an MSE of 7788.2 [m^4/h^2] and a R^2 of 0.484.

In the grid search, the ridge regression model performed best with an alpha value of 1 and with a singular value decomposition solver. For this hyperparameters, a MAE of 52.8 [m^2/h], an MSE of 8222 [m^4/h^2] and a R^2 of 0.459 were obtained. This time, the hyperparameter tuned model has the same settings as the default model, wherefore they perform equally well.

Already the default XGBoost model outperforms all other models based on the MAE. A MAE of 50.67 [m^2/h], an MSE of 7820 [m^4/h^2] and a R^2 of 0.49 is calculated. Nevertheless, the hyperparameter tuned XGBoost model shows better scoring than the default version. This is indicated with a MAE of 50.02 [m^2/h], an MSE of 7680.1 [m^4/h^2] and a R^2 of 0.50. The hyperparameters that are needed to

obtain these values are 175 for the n_estimators parameter, a learning rate of 0.1, a max_depth of 8, a gamma value of 0.25 and a lambda value of 0.3.

In Table 23, all scoring indicators and hyperparameters for the best performing combinations of all models are indicated.

Table 23: Scoring results and hyperparameter settings for the best performing versions of the MLP, the RR, the GB and the XGB model.

Model	MAE	MSE	R2	Hyperparameter	Values applied in grid search
Ridge Regression	52.8	8222	0.459	alpha	100
				solver	svd
MLP	51.1	7788.2	0.484	hidden_layer_size	(100,)
				alpha	0.0001
				activation	relu
Gradient Boosting	51.9	8076.4	0.48	n_estimators	100
				learning_rate	0.1
				max_feature	None
				max_depth	3
				min_samples_split	2
XGBoost	50.02	7680.1	0.5	n_estimators	175
				learning_rate	0.1
				max_depth	8
				gamma	0.25
				lambda	0.3

5.3.4 Testing and evaluation

To do the final testing after the hyperparameter tuning, the total area flows for PO75C and the total predicted area flows for the test data set were plotted. For better visibility, the data was aggregated into bins of 90 minutes. The outcome of this procedure can be found in Figure 51.

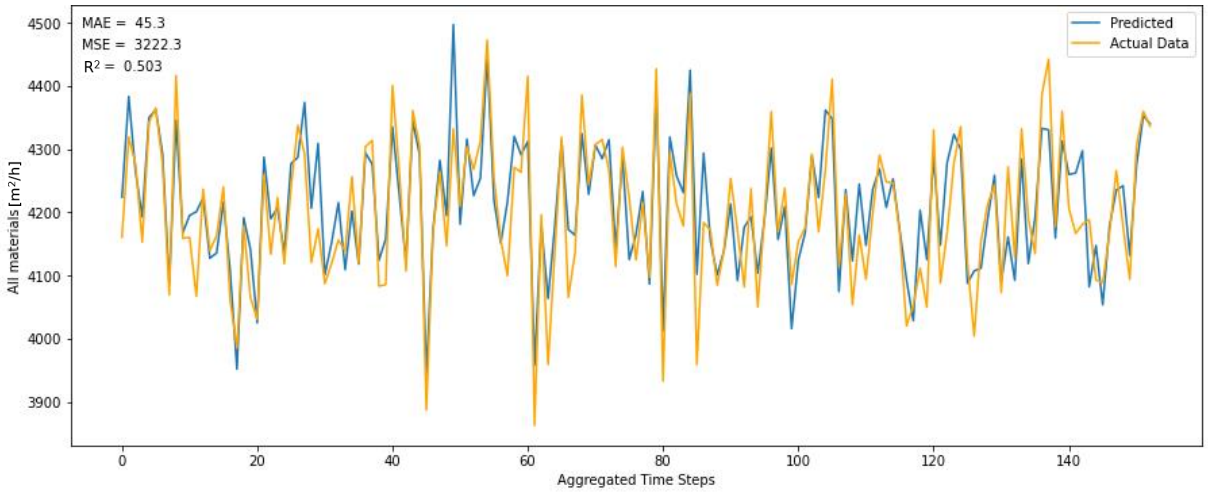

Figure 51: Predicted and measured total area flows for PO75C, aggregated for 90-minute bins with an indication of MAE, MSE and R^2 in the top left corner.

In the Figure, it can be seen that the prediction in general follows the trend of the data. The majority of the positive and negative trends is matched by the prediction, with exemplary malfunctioning at bin 11, 66 and 140. Furthermore, depressions at bin 62, 80 and 126 are underestimated. This is meant in a way, that the model detects the depression but expects a less negative peak than in the measured data. The R^2 value is not as wished with 0.503. This indicates that not all patterns in the data can be

depicted by the model. In contrary, the MAE for the aggregated data is exceptional with 45.5 [m^2/h] and also performs well for the non-aggregated case with 50.02 [m^2/h]. Compared to that, the MSE is not performing as well with 3222.3 [m^4/h^2] and 7680.1 [m^4/h^2] for the non-aggregated data. Besides the unexpected poor performance of R2 and the MSE, the small MAE gives confidence for the quality of the prediction. Transferring the obtained MAE into a percentual deviation for the aggregated data, a mean relative deviation of 1.1% is obtained.

Concluding, through the broad try out of ML models together with the hyperparameter tuning, a mean relative deviation of 1.1% for the area flow prediction from PO75 to PO75C could be reached. The MAE was identified as most important parameter during analysis, while MSE and R^2 play a secondary role. Furthermore, the best performing models were found to be RR, MLP, ET, RF, GB and XGB. Through hyperparameter tuning, XGB delivered the best results. Nevertheless, shortcomings of the grid search methodology were encountered. Therefore, with an improved hyperparameter tuning other models still could perform better.

These findings help to answer the second sub research, as they showcase an area flow prediction with the help of ML models. Furthermore, a wide array of possible improvements is indicated, which represents the potentials that lie in the methodology.

5.4 Quality prediction

As a final step, material flow prediction and area density determination will be applied together for agglomeration quality prediction in this sub chapter. This is important, as the interplay between the different developed models needs to be research. Here, it is of interest, if the intrinsic uncertainties of each model will reinforce each other or if outcomes still deliver sufficient explanatory power. The obtained insights will directly contribute to the solving of sub research question three.

5.4.1 Joint application of area density and area flow prediction

For the final quality prediction, area flow predictions from PO75 to PO75C, and area density determination for PO75C and AA106 were combined. Obtained material-specific waste flows are presented together with the belt weigher data from AA106. The described plot can be found in Figure 52.

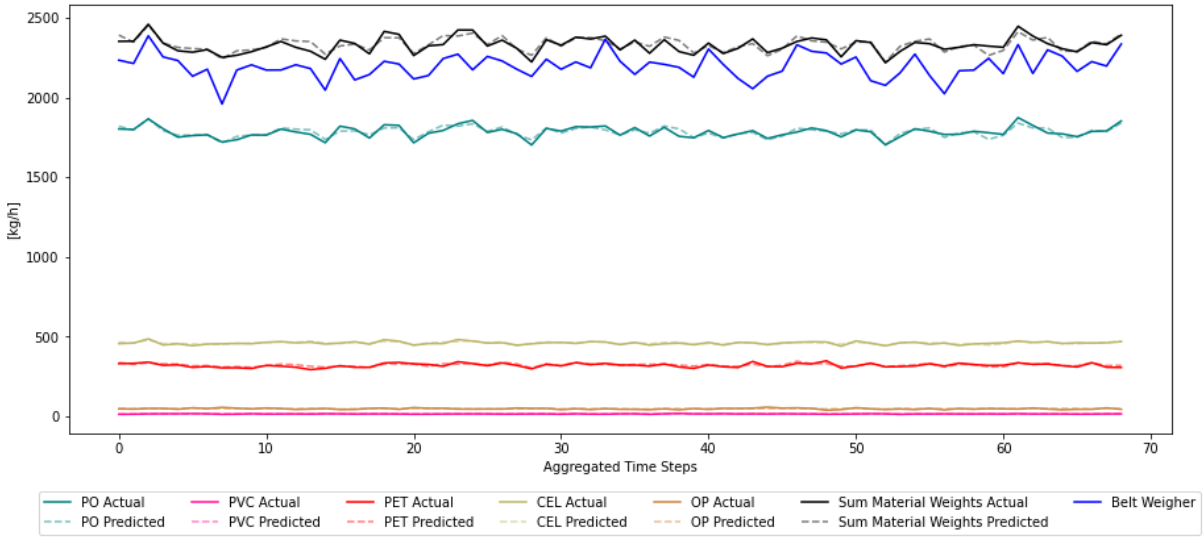

Figure 52: Predicted and measured material-specific mass flows, calculated with area densities obtained with OLS modelling from PO75C and AA106, together with belt weigher data from AA106 and the sum of specific mass flows for comparison, aggregation with 200 minutes per bin.

It can be seen, that the predicted mass flow for AA106 has an offset to the belt weigher data. A possible reason for that could be the different data pre-processing from the OLS modelling to the ML modelling. Therefore, the OLS modelling was repeated, with the data pre-processing from V9. The results can be found in Table 24.

Table 24: OLS results for area density prediction for PO75C, using data from PO75C and AA106 after similar data pre-processing to V9, grouping and drop of PET_G data.

	Regression Coefficient / Area density [kg/m²]	Standard Error [kg/m²]	95% interval
Constant	-41,656	15,662	[-72,354, -10,958]
CELLULOSICS	0,466	0,040	[0,386, 0,545]
OTHER_POLYMERS	0,376	0,025	[0,327, 0,425]
PET	3,752	0,114	[3,529, 3,976]
PO	0,464	0,008	[0,448, 0,479]
PVC	0,938	0,061	[0,818, 1,058]

Especially the constant deviates quite heavily from the value obtained with the OLS chapter pre-processing. Nevertheless, with a reduction of roughly 90% it is a favourable development. This means that the model can capture more patterns and behaviour of the data without assigning it to the constant. Furthermore, as it is not clear which materials the value of the constant belongs to, uncertainty is reduced. Similar to the previous determined area densities, bootstrapping was performed and results are presented in Table 25.

Table 25: Summary of bootstrapping results presenting the mean, median, 0.025 and 0.975 quantile for the area densities determined with PO75C and AA106 as well as V9 pre-processing.

	Mean	Median	0.025 quantlile	0.975 quantile
Constant	-41.487	-41.517	-75.141	-6.445
CELLULOSICS	0.466	0.466	0.375	0.556
OTHER_POLYMERS	0.376	0.376	0.334	0.418
PET	3.755	3.754	3.465	4.044
PO	0.463	0.463	0.445	0.482
PVC	0.939	0.939	0.829	1.052

As for the other fits, the bootstrapping results overlap well with initial OLS model results. Mean and median only show minor deviations. The strongest difference is observed for the constant with 0.17 [kg] and with 0.003 [kg/m2] for PET, representing the biggest deviation for the materials. Quantiles are wider for the constant, CELLULOSICS, PET and PO. This indicates that they have a greater variability then initially expected. Nevertheless, their ranges are still sufficiently narrow and for OTHER_POLYMERS and PVC they even outperform the initial prediction of the OLS model.

Application of the newly determined area densities can be observed in Figure 53. Here, material-specific mass flows are compiled by combining the area densities with predicted and measured area flows. Belt weigher data is provided for testing and comparison.

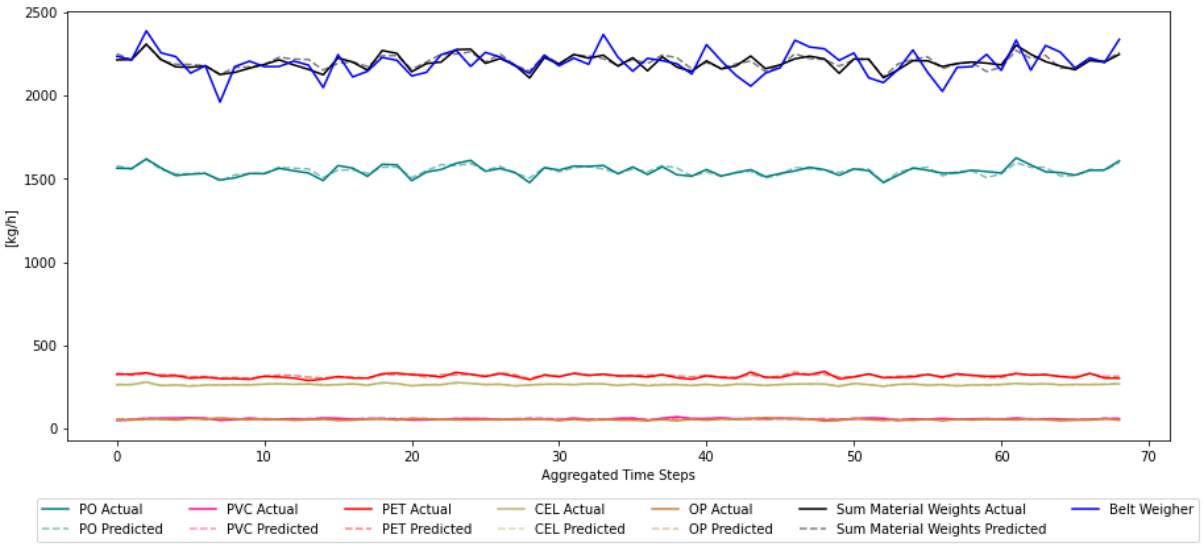

Figure 53: Predicted and measured material-specific mass flows, calculated with area densities obtained with OLS modelling and V9 pre-processing from PO75C and AA106, together with belt weigher data of AA106 and the sum of specific mass flows for comparison, aggregation with 200 minutes per bin.

This time, the sum of mass flows aligns visibly better with the belt weigher data from AA106. The offset from the previous area density fit is overcome, but the less pronounced peaks in the data remain. To gain a better understanding of possible reasons for this, the sum of material flows, their prediction and the belt weigher data is shown in a separate plot. The described plot is depicted in Figure 54.

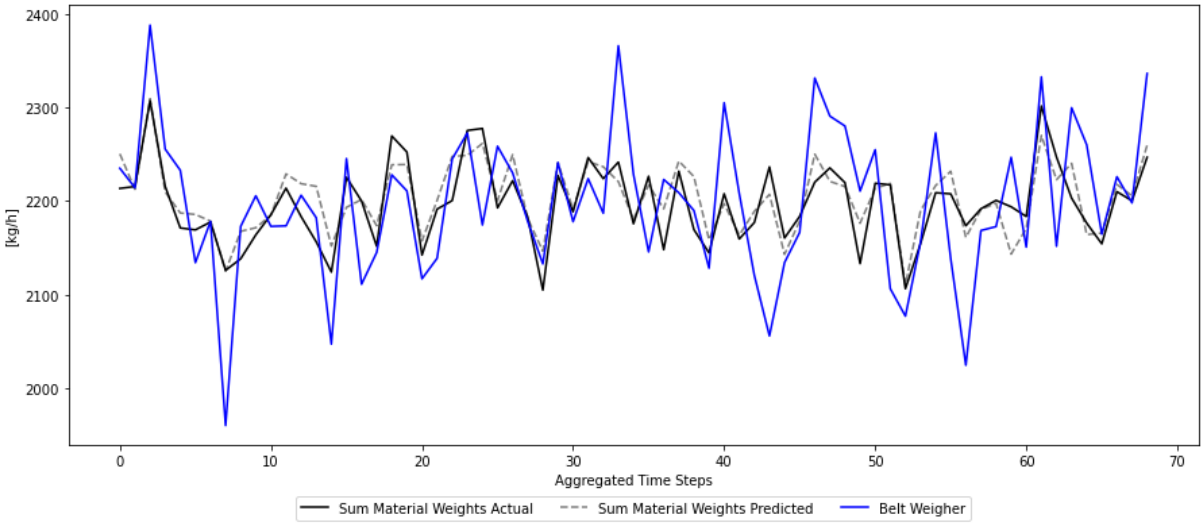

Figure 54: Belt weigher data from AA106 and the sum of mass flows based on predicted and measured area flows, and area densities obtained with OLS modelling and V9 pre-processing from PO75C and AA106, aggregation with 200 minutes per bin.

It can be seen that the general trend of the data is captured by the combination of models but that the more extreme peaks are missed. Examples for this are present at bin 7, 33, 41 and 56. For most of these peaks, the model depicts the correct sign but fails to predict the correct magnitude. This could be due to the linear nature of the area density modelling. This means that in situations where particular heavy particles pass the scanner, the same area density has to be applied as when lighter particles of the same material are present. Here, it could be interesting, if with an OLS application for shorter time periods the magnitude of these peaks would be predicted with greater accuracy.

Moving on to a more parameter driven evaluation, the MAE between the belt weigher data and the predicted sum of mass flows lies at 63.36 [kg/h] for the aggregated data. For the unaggregated data, this value increases to 651.6 [kg/h]. Especially the aggregated case is a good example for the gap in performance between R^2 and MAE. Through the slight mismatches during the observed time period the model has a low R^2 of 0.16. Nevertheless, these mismatches stay relatively close to the true values, wherefore the MAE is kept at bay.

Through the influence of the aggregation on the MAE, it can be chosen which accuracy is needed and if the prediction can still generate sufficient insights. Insights can be hampered by too high aggregation of the data. An example could be a situation where higher temporal resolution is needed. Applying the same aggregation, as for the prediction of the area flows (90 minutes), a mean relative deviation of 4% for the prediction of the belt weigher data is determined. Therefore, the error increased from 1.1% to 4% but still lies in an acceptable range.

If the area densities would be directly applied to the measured area flows instead of the predicted area flows, the relative error would lie at 3.8%. This showcases that the combined uncertainty from area flow prediction and area density determination leads to greater errors. Nevertheless, an increase of only 0.2% is a very promising result. Although this is an expected outcome, it can be seen that the uncertainty of the area densities is of greater magnitude than the uncertainty of the area flow prediction. Starting the material-specific analysis, material flows were plotted separately. Accordingly, mass flow for PO and its prediction can be found in Figure 55.

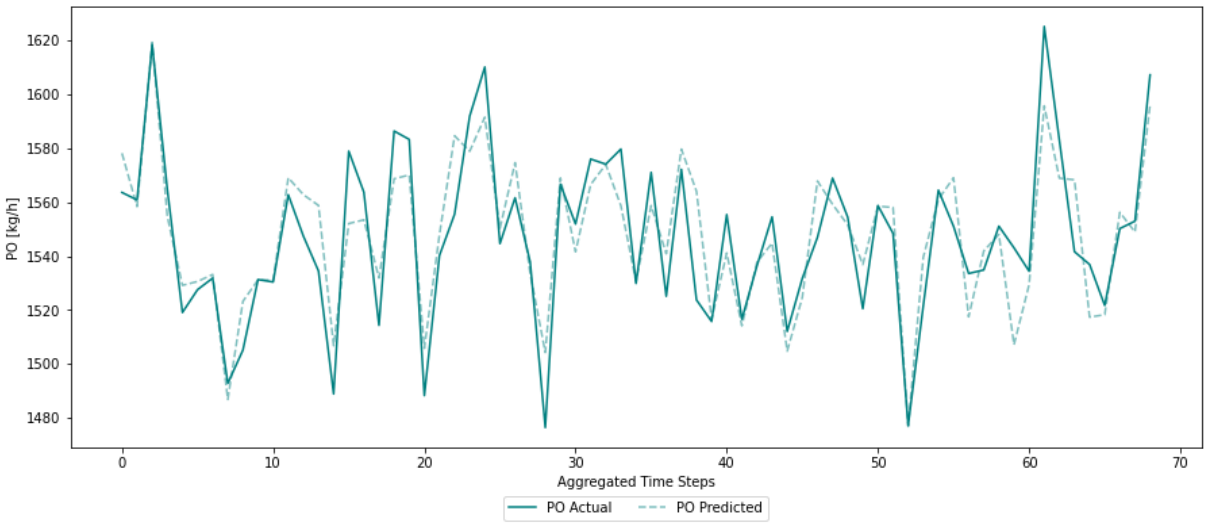

Figure 55: Mass flow for PO obtained from measured and predicted area flows of PO75C, together with determined area densities from PO75C and AA106 following V9 data pre-processing, temporal aggregation in bins of 200 minutes.

A very good fit of the mass flow for PO is observed in the plot. Almost all peaks are matched with the right magnitude and only slight variations are visible. Mismatches occur around bin 56 and 59, but the direction of the peak was predicted correctly by the model. The MAE for the aggregated data is determined with 11.6 [kg/h] and with 144.5 [kg/h] for the unaggregated data.

In the next plot, PET and CELLULOSICS are analysed. They have been grouped together, due to a convenient range for plotting and can be found in Figure 56.

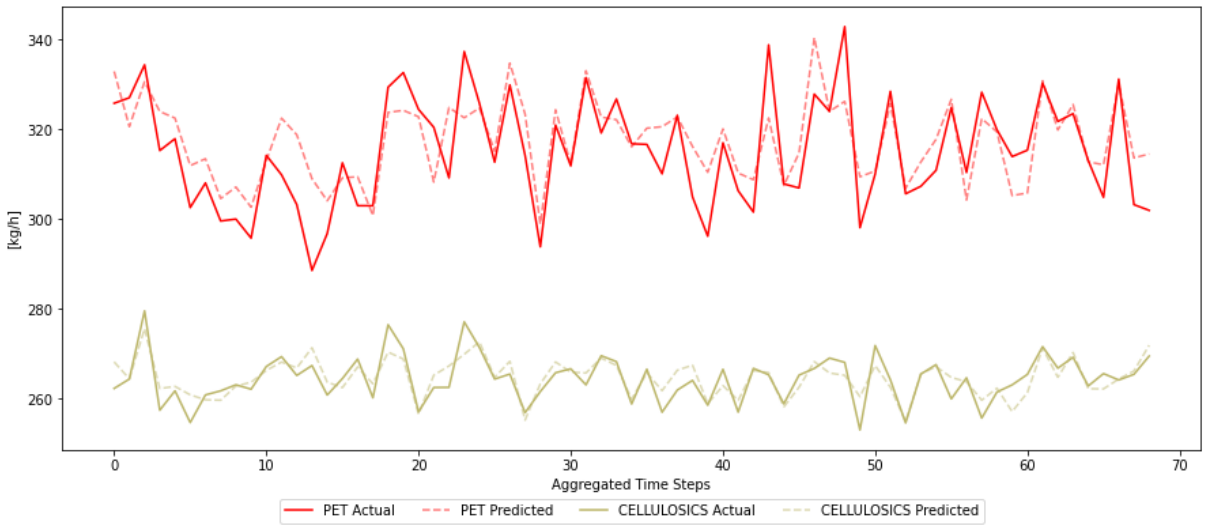

Figure 56: Mass flows for PET and CEL obtained from measured and predicted area flows of PO75C, together with determined area densities from PO75C and AA106 following V9 data pre-processing, temporal aggregation in bins of 200 minutes.

For PET, the model overestimates the majority of the data points, except for three cases at bin 17, 56 and 60. Nevertheless, the deviation is quite small, with a MAE of 6.27 [kg/h] for the aggregated data and 66.83 [kg/h] for the unaggregated data set. The CELLULOSICS show a greater variance in over-

and underestimation. This can be seen with underestimations around bin 7 and 59, as well as overestimations around bin 13 and 48. Calculating the MAE, a value of 31.14 [kg/h] and 2.34 [kg/h] is obtained for the unaggregated and aggregated case. Approaching the low magnitude materials, in Figure 57, mass flows for OTHER_POLYMERS are presented.

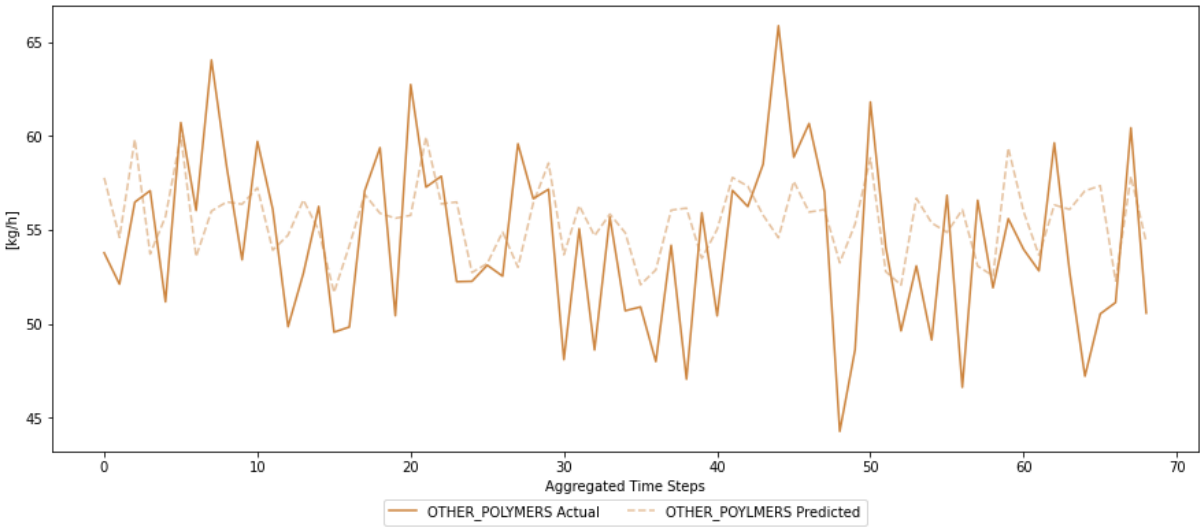

Figure 57: Mass flow for OP obtained from measured and predicted area flows of PO75C, together with determined area densities from PO75C and AA106 following V9 data pre-processing, temporal aggregation in bins of 200 minutes.

Observing the plot, the suspicion of lower accuracies and prediction abilities for lower magnitude materials is confirmed. The prediction oscillates around the mean of the mass flow with lower magnitude then the measured data. Futhermore, it fails to match the direction of the peaks frequently. Examples for this can be found around bin 13, 38, 44 and 65. Through the low occurrence, the MAE is comparable small with 3.47 [kg/h] for the aggregated data and 40.09 [kg/h] for the unaggregated data. Nevertheless, as the mean for OP is 54.36 [kg/h], this means a mean relative deviation of 73.7% for the unaggregated case. Scrutinizing the last remaining material, the mass flow for PVC is presented in Figure 58.

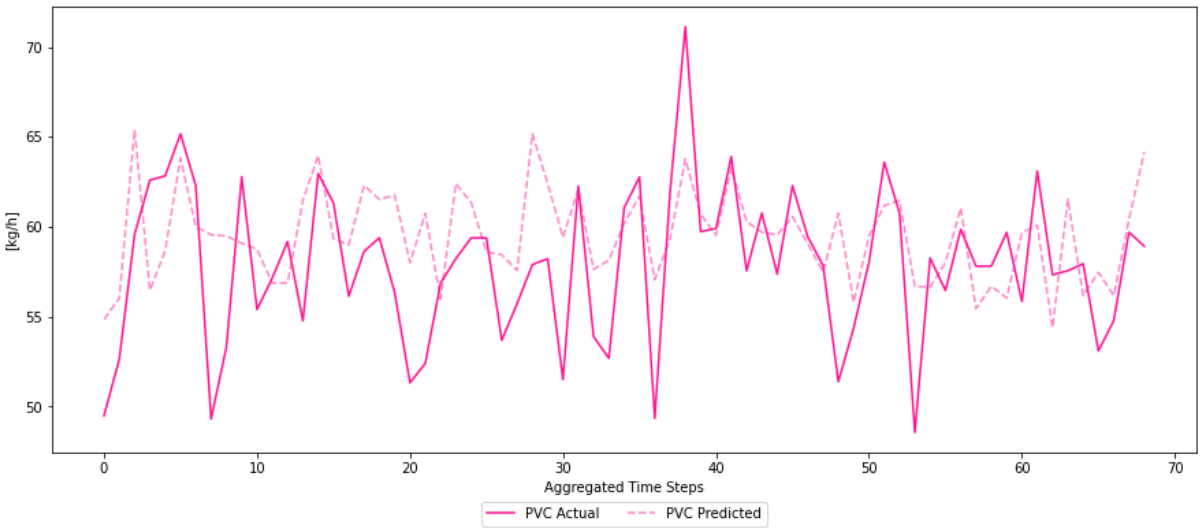

Figure 58: Mass flow for PVC obtained from measured and predicted area flows of PO75C, together with determined area densities from PO75C and AA106 following V9 data pre-processing, temporal aggregation in bins of 200 minutes.

It can be seen that the trend from the OTHER_POLYMERS plot can be confirmed with PVC. Direction and magnitude of the prediction are mostly wrong and tendency for overestimation is seen between bin 15 and 33. On the positive side, no consistent trend for underestimation is present. This is favourable, as for the quality assessment of the agglomeration product it is of importance that PVC is not underestimated. Through the tendency of overestimation, a surprisingly increased PVC occurrence during real world testing of the quality becomes less probable. The MAE for PVC is determined with 3.32 [kg/h], for the aggregated data, and with 27.55 [kg/h], for the unaggregated data. This signifies an MRE of 47.6%.

As a last step, the percentual weight share of each material was determined and plotted over time. Result are presented in Figure 59.

Figure 59: Weight shares on PO75C resulting from measured and predicted area flows, calculated with area densities obtained with OLS modelling and V9 pre-processing from PO75C and AA106, aggregation with 30 minutes per bin.

Relatively stable percentual mass shares can be observed, which oscillate around a common mean in small magnitudes. Compared to the mass flow plots, a way smaller aggregation is chosen with 30-minute time steps instead of 200 minutes per bin. Still peaks and changes in trend are visible, which speaks for the stability of the data. PO has a mean weight percentage of 68.9%, PET was calculated with 14.1%, CELLULOSICS were determined with 11.8%, while OTHER_POLYMERS have a mean share of 2.5% and PVC has a mean percentage of 2.7%.

5.4.2 Separation efficiency

To finalize the quality prediction of the agglomeration product, the separation efficiency of PO75C has to be estimated. As PO75C is the last NIR-scanner before the dispatch of the product, no material characterization after the separation work of PO75C is available. Therefore, conclusions have to be drawn from the separation step between PO75 and PO75C.

For this, the assumption is made that targeted and untargeted materials behave similar for PO75 and PO75C. This means that the separation efficiency from PO75 to PO75C will be transferred to the separation step from PO75C to the final product. For the separation step from PO75 to PO75C,

PE_FILM, PE_RIGID, PP and PP_FILM (later grouped as PO) are the targeted materials. Observing the separation step from PO75C to the final product, PVC and PAPER are ejected and the non-targeted particles result in the agglomeration product. To transfer the separation efficiency, the share of material that was transferred from the first to the second stage has to be determined. Furthermore, information is needed how much non-targeted material was removed together with the targeted material.

To determine the separation efficiencies, area densities for PO75 were ascertained with AA101 belt weigher data and V9 data pre-processing. Results are presented in Appendix 8. This was done to match the data pre-processing for the area densities from PO75C and to calculate the separation efficiency through the obtained mass flows. PO was calculated with a mean mass flow of 1294.3 [kg/h], for PO75, and with 1551.28 [kg/h], for PO75C.

As the material cannot increase from one scanner to the next, this resembles an impossible result. This could be due to measurement errors, the general uncertainty of the scanners, changed categorization based on changed target materials or because of the uncertainty from the area density determination. Especially the area densities for PO75 have a high constant of 602.3 [kg/h], compared to -41.49 [kg/h] for PO75C. This results in a range of 643.8 [kg/h] of unexplained mass flow based on area densities, which could explain the unrealistic increase of PO.

To decrease the uncertainty for separation efficiency determination, it was fallen back on the comparison of area flows from PO75 to PO75C. As separation efficiency is material-specific, it turned out that no broader context of other materials present on the scanner is needed. Therefore, the conversion into mass flows is not necessary. Through determination of separation efficiencies before the conversion into mass flows, uncertainty of the area densities is added later. Therefore, deducted separation efficiencies are more accurate. To use this advantage, the mean PO content on PO75 was determined with 3434.6 [m^2/h] and a mean PO content of 3342.94 [m^2/h] was obtained for PO75C. This indicates a separation efficiency of 97.3%. In a next step, the percentual area share of PO on PO75C was determined with 79.5%. This means that, with the targeted material, an additional amount of material that has an area of 25.8% of the moved targeted material is transferred.

Assuming that these figures are applicable for the separation step from PO75C to the final product, this means that 97.3% of the PAPER and PVC area flow will be removed. Furthermore, additional material, of the area of 25.8% of the removed PAPER and PVC, will be transferred with it. Here, the assumption is made, that these 25.8% are spread over the remaining materials according to their share of area flow. Inconveniently, the PAPER category was grouped together with BC and BOARD_CT for area density determination to address multicollinearity. This is a problem, as in the next step area densities need to be applied, but the material composition of the material grouping is altered.

To mitigate this alteration, area densities were determined again for PO75C together with V9 data pre-processing and the goal of keeping PAPER a distinct category. Complete results can be found in Appendix 9. Checking for multicollinearity, correlation between BC and PO was considered too high with 0.83. Furthermore, 0.025 and 0.975 quantiles are wider compared to CELLULOSICS if PAPER and BC are kept separate. Therefore, area densities with the grouping of BC and PAPER as CELLULOSICS were kept. This decision was strengthened by the fact that area densities for the respective categories were closely together, with 0.452 [kg/m^2] for BC, 0.466 [kg/m^2] for BOARD_CT and 0.483 [kg/m^2] for PAPER. Here, it is favourable that PAPER has the highest area density. Through the removal of PAPER area flows, more dense particles are removed in reality, but due to the

application of the CELLULOSICS area density, less material is removed in the model. Therefore, results are on the safe side of the estimation, as in reality product purity should be slightly better. In Figure 60, an application of the determined separation efficiencies to the material-specific mass flows on PO75C can be found.

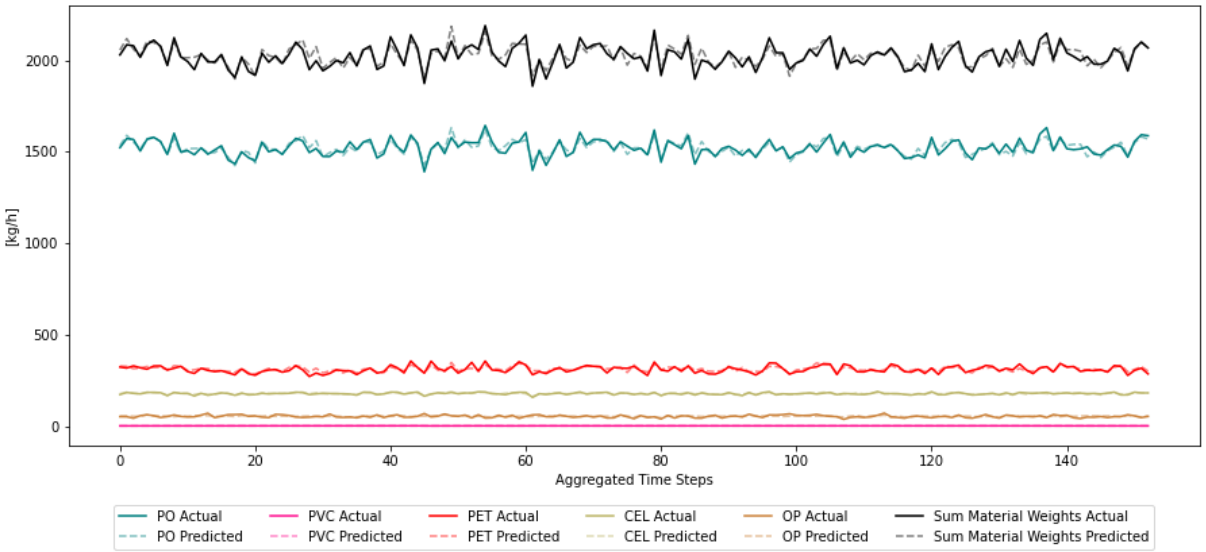

Figure 60: Material-specific mass flows of the agglomeration product resulting from predicted and measured area flows, calculated with area densities, obtained with OLS modelling and V9 pre-processing from PO75C and AA106, together with application of separation efficiencies as well as belt weigher data from AA106 for comparison. Aggregation with 90 minutes per bin.

Compared to Figure 53, PVC and CELLULOSICS show the greatest deviation. Before, CELLULOSICS were on a similar magnitude to PET, while now they are clearly separated. CELLULOSICS are observed on a lower occurrence level, with a mean of 178.13 [kg/h] for the predicted value. PVC also changed in magnitude and is separately visible from OP in this version of the plot. The new mean of the PVC category is 1.6 [kg/h] for the predicted value.

5.4.3 Prediction results

Through the joint application of area flow prediction, area density determination and separation efficiency approximation, the quality of the agglomeration product can finally be predicted. Results for the test data set can be observed in Figure 61.

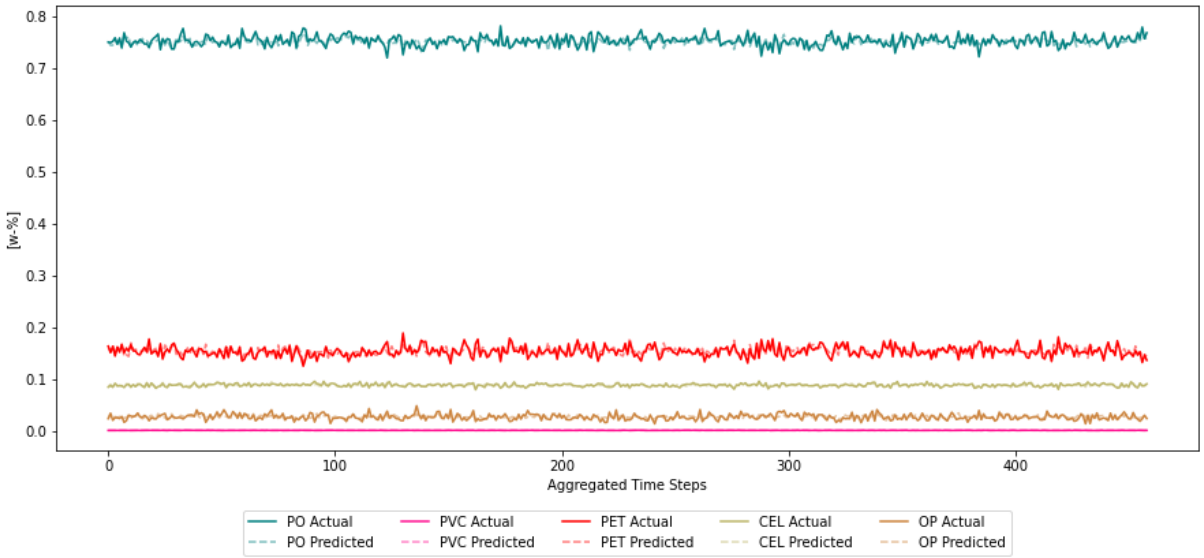

Figure 61: Percentual weight shares for measured and predicted area flows after application of separation efficiencies, calculated with area densities obtained with OLS modelling and V9 pre-processing from PO75C and AA106, aggregation with 30 minutes per bin.

Similar to the previous weight percentage plot, a very stable percentual weight share can be observed. This stability facilitates to aggregate the time steps to only 30 minutes per bin, while results are still visibly assessible. Mean weight shares for PO of 73.7% and 73.6%, for the measured and predicted area flows, are obtained. Maximum values are detected with 76.4% and 75.1%, for measured and predicted area flows, while minimum values lie at 70.5% and 71.8% respectively. The MAE, from predicted to measured area flows for PO, lies at 0.34%. For PVC, mean weight shares are determined with 0.08%, for both area flow determination methods. Maximum and minimum values lie at 0.12% and 0.11%, as well as 0.05% and 0.06%, for measured and predicted area flows. Comparing the weight shares resulting from predicted and measured area flows, it is seen, that percentages obtained from the predicted area flows underestimate the percentage in both directions. This can be explained by the same trend, that was observed for area flow prediction and mass flow determination. As the final quality prediction is composed of these two modelling parts, it resembles an expected finding.

Circling back to the quality requirements introduced at the beginning of this work, the quality requirement for PVC can be met with high security. A mass percentage below 1.4% of PVC is needed and would be still achieved with separation efficiencies of down to 69.5%. On the other end, PO quality criteria with a minimum of 75% weight share are not met. Nevertheless, this could change with a higher separation of PET or CELLULOSICS, which could be targeted together with PVC and PAPER on PO75C. PET has a mean weight share of 15% and 15.1%, and CELLULOSICS are present with 8.6%, for measured and predicted mass flows.

6 Conclusion and recommendations

The motivation of this research was to help with the implementation of improved plastic waste separation, to unlock environmental and economic advantages. To achieve this, real-time machine setting optimization in waste separation plants was identified as one of the most promising approaches. This type of optimization is enabled by dynamic separation efficiency modelling and real-time waste stream characterization. As extensive waste stream characterization with sensors is economically prohibited, the use of data from NIR separation units, that are already present in the plant, is sought after.

To explore the joint application of these concepts, the prediction of the HQ agglomeration line product in a plastic waste separation plant in Scandinavia was analysed. In the agglomeration line, two NIR-separators are connected in series to ensure high-quality separation of the material. The input to both NIR-scanners is measured by belt weighers. Through the belt weighers and NIR-scanners, total mass flow and material-specific area flow data is available. In total, 4-month worth of data was provided. The specific goal of the research was to predict the agglomeration line product quality, after the material has been processed by both separators. Here, data from the first NIR-scanner serves as the basis for the prediction. Quality requirements of the product are defined by weight shares of PO and PVC. PO is the target material with a minimum presence of 75 w-%, while PVC resembles the most important contaminant with a maximum occurrence of 1.4 w-%.

To model the quality prediction, material-specific mass flow data needs to be obtained. As only material-specific area flows are available, this can be achieved by the determination of area densities. When area densities are obtained, the prediction of area flows from the first to the second NIR-scanner and from the second NIR-scanner to the product is needed. Alternatively, separation efficiencies can be used. Through a joint application of these concepts, the agglomeration line product quality can be predicted in weight percent.

To guide the modelling of the agglomeration line product quality, the following research question was compiled:

How can the quality of the agglomeration line product in a plastic waste separation plant be predicted based on NIR-scanner and belt weigher information through a data driven approach?

To split up the research question in workable sections, three sub research question were formulated:

- *What correlations and relationships exist in the data?*
- *How can the area density, the area flow prediction and the separation efficiency be modelled?*
- *How does a joint application of the developed concepts perform for quality prediction of the agglomeration line product?*

6.1 Summary and conclusions

In this section, the research question and sub research questions are answered. Furthermore, the derivation of the results is summarized.

Starting with the first sub research question, data exploration for belt weigher data, NIR-scanner area flow data and NIR-scanner belt occupation data was conducted. This was needed, as prediction quality is highly dependent on the quality of the input data. Therefore, pitfalls in the data and remedies for these pitfalls have to be identified. Furthermore, present correlation in the data delivers insight about potential modelling approaches and how the data can be made usable for the goal of this research.

For the belt weigher data, a tendency was found that values close to zero instead of zero are measured. Through this effect, the cleaning of the data set from empty data points is hampered. To avoid this, different thresholds for data exclusion were tested. Best results were obtained with a threshold of 0.3 [t/h]. After this clean-up, a correlation of 0.867 between belt weigher AA101 and AA106 was obtained. Furthermore, it could be ascertained that the data is free from temporal correlation patterns.

Regarding area flows, a similar behaviour to the belt weighers was found. It could be shown, that the NIR-scanners tend to measure values close to zero instead of zero. This happens when either the belt is empty or the scanners are malfunctioning. Furthermore, it was observed that this effect applies for all materials together, but also for single materials. During the search for temporal correlation patterns, it was found, that correlation drops of the materials do not overlap. Furthermore, for low area flow material, lower correlation and increased erroneous behaviour was observed. These findings strengthen the hypothesis of material-specific malfunctions in detection. Furthermore, a state of the machines was encountered, where the total detected area drops to a threshold and then oscillates around this value with small magnitudes. In this state of stagnation, different materials are either measured with constant values or increasing or declining behaviour in a balancing fashion. This phenomenon was observed for several materials on NIR-scanner PO75C. Values for these stagnations were total area flows around 500 [m^2/h] or 1000 [m^2/h]. To tackle the encountered problem, thresholds for data inclusion, of total area flows on the NIR-scanners PO75 and PO75C, were defined. This was done with 1000 [m^2/h] and 12,500 [m^2/h] for PO75 and 1250 [m^2/h] and 10,000 [m^2/h] for PO75C. The upper thresholds were motivated from distribution analysis of the total area flows. They were implemented to exclude high magnitude outliers. Regarding correlation, the NIR-scanners correlate reasonably after application of the thresholds with a r value of 0.829. PE_FILM and PE_RIGID were the highest correlating materials with r values of 0.9 and 0.89. OTER_POLYMERS and PS were the materials with the lowest correlations with r values of 0.07 and 0.13. Lastly it could be determined, that the data is free from temporal correlation patterns.

For eased exploration of belt occupation data, belt occupation was divided in 5 categories. Next, the conveyor belt was split into 70 compartments over its width and occurrences of the categories were counted. The 50-100 [m^2/h] category was most appearing (2.14 million) followed by the 0-50 [m^2/h] category (2.08 million). Lower occurrences were detected for the 100-200 [m^2/h] category (1.17 million), while the 200-300 [m^2/h] category (11,704) and the >300 [m^2/h] category (1,144) were rarely present. Regarding correlation, the 0-50 [m^2/h] category showed a negative trend with high magnitude area flows. In contrast, the 100-200 [m^2/h] category showed a positive trend towards it. In general, the 100-200 [m^2/h] category was identified as the governing category for high magnitude area flows. The 50-100 [m^2/h] category and the 200-300 [m^2/h], as well as the >300 [m^2/h] categories showed no

significant linear trend during analysis. As a great variance of non-linear patterns was visible, it was decided to integrate the data into a ML approach.

Although not part of the official data exploration, insights about the joint use of belt weigher and area flow data were gained during area density modelling. Here, a clear temporal correlation pattern was encountered. Correlation first oscillates around r values of 0, then they elevate to r values of 0.8 with a drop to r values of 0.6 for a short period of time. No clear explanation, apart from machine malfunctions, could be found. To counteract this effect, all data below a r value of 0.7 was excluded. After removal, a correlation of 0.827 between belt weigher AA101 and NIR-scanner PO75, and a correlation of 0.822 between belt weigher AA101 and NIR-scanner PO75 was obtained.

The conducted data exploration answers the first sub research question *"What correlations and relationships exist in the data?"* in the following way: Belt weigher data correlates well with each other. Problems due to a tendency of measuring values close to zero instead of zero arise. Excluding values below 0.3 [t/h] was found to handle this issue reasonably well. NIR-scanner data offers a wide range of inconsistencies. Hot spots are total area flows below 1000 [m2/h] and low magnitude area flow materials. Apart from that, correlation is high. Encountered problems can be fixed by applying thresholds of 1000 [m^2/h] and 12,500 [m^2/h], for PO75, and 1250 [m^2/h] and 10,000 [m^2/h], for PO75C. Belt occupation data delivers low explanatory power by linear correlation. The 100-200 [m^2/h] category is governing for the correlation that was found and an array of non-linear patterns is present. Therefore, the information hidden in the data could be leveraged through a ML approach. Joint belt weigher and area flow data application uncovered an unexpected temporal correlation pattern. A potential remedy is to remove time periods below a defined correlation threshold.

Starting into the quality prediction modelling, first the area densities had to be determined. This was done to enable the conversion of material-specific area flows into material-specific mass flows. Material-specific mass flows are needed, as quality is determined in weight percent.

To do this, OLS methodology, as most suitable linear regression technique, was applied. Multicollinearity was encountered and treated by grouping of the materials. Final material categories were CELLULOSICS, OTHER_POLYMERS, PET, PO and PVC. The effects of multicollinearity were successfully mitigated through this approach. During the course of the thesis, the insight was gained, that determined area densities are sensitive to different data pre-processing. This showcases the heavy influence of data input on the modelling outcome and highlights the need of similar data pre-processing during joint application of modelling approaches. Furthermore, it was proven that area densities are not generalizable from NIR-scanner to NIR-scanner. This is due to the material composition change during sorting. Through bootstrapping, greater trust in the confidence intervals for the area densities could be obtained. Calculated 0.025 and 0.975 quantiles match the confidence intervals for all area densities, with slightly increased magnitude. The area densities that were obtained for final application showed standard errors below 0.01 [kg/m^2] and confidence interval spreads under 0.3 [kg/m^2]. During application on the test data set, a MAE of 141 [kg/h] for an aggregation of 30 minutes was observed. This resembles an MRE of 3%.

In a next step, the material-specific area flow from the first to the second scanner was modelled. This was necessary, as the final quality prediction is intended to solely rely on data from the first NIR-scanner. While the area density was modelled to have the right unit for the prediction result, the area flow prediction resembles the first prediction component of the research.

For area flow prediction, nine different data subsets and data pre-processing were tested. This was inspired by the upfront conducted data exploration. During a broad model try-out, the area flows on NIR-scanner PO75C were predicted based on data from NIR-scanner PO75, belt weigher AA101 and belt weigher AA106. This was done with the following models: Decision Tree, Extra Tree, Random Forest, Gradient Boosting, K-Nearest Neighbours, Bagging Regressor, Ridge Regression, Elastic-Net Regression and Multi-Layer Perceptron. During analysis, it was noted that versions with thresholds for total area flow performed better regarding MAE and MSE. Versions without these thresholds showed better results for R^2. For model selection and further analysis, it was decided to focus on MAE over MSE and R^2. This was done, as smaller general errors were preferred over better detected patterns and improved outlier handling. The best performing data base and data pre-processing version was picked with V9. This includes a drop of NaN and zeros, integration of two additional material categories on PO75, data from AA101 and AA106, belt occupation counts and thresholds for total area flows on PO75 and PO75C. Apart from that, Random Forest, Gradient Boosting, Ridge Regression, Extra Tree and Multilayer Perceptron were the best performing models during the try-out. Further selection was based on the representation of different model architectures and expected improvement during hyperparameter tuning. Accordingly, Ridge Regression, Multilayer Perceptron and Gradient Boosting were selected. Additionally, XGBoost was added to the hyperparameter tuning, as it resembles an improved version of Gradient Boosting. During hyperparameter tuning, XGBoost outperformed the other models. This was achieved with hyperparameter values of 175 for n_estimators, 0.1 for the learning rate, 8 as max depth, 0.25 for gamma and 0.3 for lambda. Finally, a MAE of 50.02 [m^2/h], an MRE of 1.1%, an MSE of 7680.1 [m^4/h^2] and a R^2 value of 0.5 was obtained.

The last needed information, to assemble the quality prediction of the agglomeration line product, is the separation efficiency from the second NIR-scanner (PO75C) to the final product. Together with the area densities and the area flow prediction, from the first to the second NIR-scanner, mass-based material composition on the second NIR-scanner can be determined. Therefore, only the separation efficiency from PO75C to the agglomeration product is missing to achieve agglomeration product quality prediction.

For separation efficiency determination from PO75C to the final agglomeration product, no data driven modelling approach could be identified. This was due to missing validation data. To still reach a result, the assumption was made that separation efficiencies from PO75 to PO75C are generalizable to the step from PO75C to the product. For targeted material, a separation efficiency of 97.3% was obtained. Additionally, the amount of non-targeted material, ejected along with the targeted material, was estimated. For this, a material flow, of the size of 25.8% of the successfully ejected target material, was determined.

The summarized modelling section answers the second sub research question *"How can the area density, the area flow prediction and the separation efficiency be modelled?"* in the following way: Area densities can be modelled by an OLS approach. Multicollinearity is successfully treated through material grouping. Together with the application of bootstrapping, greater trust in the confidence intervals and calculated area densities is obtained. The procedure enables the determination of area densities with an MRE of 3%. Area flow prediction can be obtained through a ML approach. With the help of a broad try-out of models, data subsets and data pre-processing, their best combination can be determined. Together with hyperparameter tuning, a MAE of 50.02 [m^2/h], an MRE of 1.1%, an MSE of 7680.1 [m^4/h^2] and a R^2 value of 0.5 could be obtained for the XGBoost model. For the separation

efficiency from PO75C to the final product, no data driven modelling approach could be identified. Nevertheless, through the assumption of generalizability of separation efficiencies from PO75 to PO75C to the last separation step, separation efficiencies could be approximated. Therefore, a separation efficiency of 97.3% and 25.8% additional material transferal could be estimated.

To conduct the final quality prediction, the determined area densities, area flow predictions and separation efficiencies were applied together. First, the area flow prediction was joined with the area densities to determine the mass flows on PO75C. It was found, that the OLS model has to be trained with the same data pre-processing as the ML model to obtain the best results. Testing was done with the test data set and AA106 belt weigher data. For a 90-minute aggregation, an MRE of 4% was found, for application of predicted area flows, and an MRE of 3.8% was obtained, for application of measured area flows. Therefore, through the use of predicted area flows the MRE only increased by 0.2%. Through application of separation efficiencies, conversion into mass flows and calculation of their mass share, the final quality prediction was determined. For a 30-minute time step aggregation, PO showed a mean weight share of 73.7%, for application of measured area flows, and 73.6%, for application of predicted area flows. Maximum and minimum values were 76.4% and 75.1%, as well as 70.5% and 71.8%, for application of measured and predicted area flows. The MAE from measured to predicted area flows was obtained with 0.36%. For PVC, a mean weight share of 0.08%, for measured and predicted area flows, was determined. Maximum and minimum values were 0.12% and 0.11%, as well 0.05% and 0.06%, respectively. Furthermore, a MAE of 0.007% was calculated. With the described values, the quality criteria for PO were not met, but the quality criteria for PVC was held by a fair margin.

With the obtained results, the third sub research question *"How does a joint application of the developed concepts perform for quality prediction of the agglomeration line product?"* can be answered as follows: If the assumption of transferability of separation efficiencies from PO75 to PO75C to the final separation step holds, the prediction of PO has an expected MAE of 0.36%. For PVC a MAE of 0.007% is anticipated. Regarding the interplay of the models, greater uncertainty results out of the area density determination. This is based on the mass flow prediction on PO75C. Here, the area density contributed 3.8% to the MRE and the area flow prediction only was responsible for 0.2% of the MRE indicator.

Joining the obtained information, the main research question *"How can the quality of the agglomeration line product in a plastic waste separation plant be predicted based on NIR-scanner and belt weigher information through a data driven approach?"* is resolved in the subsequent way: The quality of the agglomeration line product, in a plastic waste separation plant, can be predicted using multiple modelling building blocks. An OLS model is applied to determine area densities. Subsequently, an XGBoost model is used to predict area flow. Finally, separation efficiency is transferred from a representative separation unit to complete the prediction process. Through this, MAE for weight share prediction of 0.36%, for PO, and 0.007%, for PVC, for a 30-minute time step aggregation are obtained.

6.2 Recommendations

In this section, recommendations that are derived from the findings of the study will be presented.

During data exploration it was seen, that especially the use of NIR-scanner data, for waste stream characterization and quality prediction, should be done with great care. Therefore, it is highly recommended to explore all data thoroughly. This is particularly the case before model building, to avoid incorporation of inconsistency through the data. Especially, the stagnation of total area flow around 1000 [m^2/h] for the NIR-scanners show an interesting example of inconsistencies. Without data exploration, this erroneous data would have been incorporated into the model. Therefore, all input data should be scrutinized for inconsistencies and remedies for encountered problems should be thought of.

Working with the data used in this thesis, upper and lower total area flow thresholds should be applied. These are 1000 [m^2/h] and 12,500 [m^2/h], for PO75, and 1250 [m^2/h] and 10,000 [m^2/h], for PO75C. To handle inconsistencies of belt weigher data, a threshold for data inclusion of 0.3 [t/h] is imperative. Furthermore, it is recommended to compute correlation over time, to check for temporal correlation patterns. If temporal correlation patterns are encountered, the exclusions of data points with low correlation should be thought of. Here, the threshold is case dependent and should be determined based on the needs of the given modelling task.

Apart from this, it is recommended, to seek better understanding of the inner workings of the NIR-separators. For this, contact with TOMRA should be sought to discuss and understand the encountered inconsistencies. Here, especially the stagnations states of the machines and the reduced accuracy of low area flow magnitude materials is of interest.

Delving into specific model components, an improved separation efficiency determination from PO75C to the agglomeration product should be found. This is of importance, if the developed methodology is applied for quality prediction in a real-world separation process. Here, the use of a classification unit is recommended. If economic constraints prevent this, probing or expert knowledge can be potential remedies. Furthermore, using a more representative unit pair for transferring separation efficiencies to the last separation step, is recommended to gain greater trust in the prediction. Additionally, more direct validation methods for the remaining modelling components should be developed. By now, only the area flow prediction is directly validated through area flow data from the second NIR-scanner. If results of this work can be confirmed, a way must be found to meet the quality requirements for PO. This can be done by targeting CELLULOSICS or PET next to PAPER and PVC on PO75C.

An interesting point, that was revealed during the writing of this work, is the influence of temporal aggregation on the MAE and the MRE. Here, it is recommended to find the needed temporal aggregation for plant optimization and focus further model building on this time scope. Apart from that, the ideal combination of pre-processing approaches for the interplay of models should be found. The development of a respective methodology is recommended. Furthermore, the amount of data, used for area density determination, should be reviewed. The OLS model searches for the area densities that give the smallest possible error over the given time frame. As the used data has a time span of 4 months and waste composition is constantly changing, uncertainty could become smaller with the use of smaller time frames. Therefore, it is recommended to research this part before further applying the developed area density determination. Switching the focus to the ML modelling, an improved hyperparameter

tuning methodology should be implemented. This is due to the fact, that the implemented hyperparameter tuning in this study showed only minor effects, but bigger potentials are expected. Here, either a more extensive use of grid search, the combined use of grid search and random search or implementation of more advanced hyperparameter tuning methods should be applied.

Thinking of a more widespread application of the showcased concepts, a use of all available characterization information is recommended. For this, each belt weigher NIR-scanner pair in the plant should be used for area density determination. Afterwards, area densities should be analysed, aiming for insight about needed air nozzle pressure for ejection. If this information can be retrieved, fine tuning of air nozzle pressure is recommended to unlock economic and environmental advantages. Furthermore, through the obtained area densities, material-specific mass flows determination should be implemented for each belt weigher NIR-scanner pair. Through this, mass-based material composition changes, from NIR-scanner belt weigher pair to NIR-scanner belt weigher pair, can be revealed. With this, information about separation efficiencies gets unlocked and plant optimization should be put into action. Here, changes in machine settings can be analysed more precisely and their effect can be evaluated. Furthermore, prediction elements of the showcased modelling should be added to the described approach. Through this, simulation, of the effects of machine setting changes, can be developed. Apart from that, if waste stream characterization and prediction is implemented in detailed temporal granularity, the plant can be optimized in real-time. This would enable the sorting of each waste input in an optimal way and therewith unlock further economic and environmental advantages. Therefore, setting the described implementation as a long-term goal is highly recommended.

6.3 Limitations

Limitations arise from different parts of this work. These can be split up into data exploration limitations, model building and model capability limitations as well as limitations resulting from the higher-level approach of this research.

Starting off with the data exploration limitations, it is not guaranteed that the data is sufficiently explored. This means that correlations and relationships, that would enable improved modelling, could remain unrevealed. On the contrary, crucial inconsistencies, that hamper the explanatory power of the developed model building blocks, could have been missed. Apart from that, not all available data from the TOMRA Insight portal was used. Therefore, useful relationships in the data could remain unexplored. Strong limitations were encountered for NIR-scanner data. This is mostly due to the described stagnation states of the machines and the measurements of values close to zero instead of zero. Furthermore, the explanatory power for lower magnitude material is limited and the results cannot be used with the same confidence as for high magnitude materials

Delving into the modelling part, the limitations of the area density determination must be understood. The area densities can only be validated indirectly. Through a probing and characterization campaign this could be done in the future, but by now the area densities were only validated indirectly. This means, that the material-specific mass flows were summed up and tested against belt weigher data. Here, the resulting error was evaluated. Bootstrapping and confidence intervals were applied, but it is important to understand that they do not come with the same validity, as validation data would provide. The biggest limitation of this study is resembled by the separation efficiency determination of the last separation step. By now, there is no data to validate it, also not indirectly. Furthermore, no statistical

methods to enhance trust in the obtained results were conducted. The only source of confidence is that the separation efficiencies were transferred from a previous unit of the same plant. Therefore, the determination of the separation efficiency from the second NIR-scanner to the agglomeration product, should be solely seen as a measure to answer the initial research question, but not as a recommendation for implementation.

Regarding the area flow prediction from PO75 to PO75C, limitations arise from the lower prediction accuracy for lower magnitude area flow materials. With respect to hyperparameter tuning, grid search does not guarantee to find the optimal hyperparameters. Therefore, limitations to the model capabilities occur. Further limitations arise due to the sole focus on MAE, as other indicators where not considered during final ML model building. Lastly all trained models can only learn from data that they have seen. This signifies a limitation to past waste compositions. Therefore, if new materials are introduced into the waste stream, other sorting behaviour could occur, which limits the explanatory power of the model.

From a methodological viewpoint, no framework was set up to identify modelling approaches for the goal of the research. Therefore, this research could be limited by better suited methodologies that were missed. Furthermore, the final goal of achieving a real-time waste sorting plant optimization was identified but better suited approached for improved plastic waste separation could exist. Zooming out even further, other approaches for waste management improvement, that yield greater economic and environmental advantages than improved plastic waste separation, could be worth investing time and effort into. Putting it differently, limitations could arise, through more effective higher-level methodological approaches that were not considered.

6.4 Future research

Future research is recommended for areas of this work that come with greatest reduction in uncertainty. Furthermore, research that enables the widespread use of the explored concepts as well as methodologies that improve the approach of this work should be explored.

Regarding uncertainty reduction, the separation efficiency determination, from PO75C to the final product, should be improved. This can be done by conducting research on how validation data for this step can be retrieved in an economically sound way. If this is not possible, methodologies on how to approximate the sought information, while minimizing uncertainty, should be explored. Potential research approaches could be the transferral of separation efficiencies from other units, including a proof of their similarity, or application of statistical methods, to gain greater trust in the retrieved values. Regarding area densities, it should be researched how they can be validated directly, instead of indirectly, in an economically sound way. Apart from that, exploration of the influence of data set size and data pre-processing, for the area density modelling is of interest. It is suspected that with smaller data set sizes the area density determination could become more accurate. This would also decrease uncertainty of material-specific mass streams and uncertainty of separation efficiency determination between belt weigher NIR-scanner pairs, which shows the advantages of respective research. To minimise uncertainty of material-specific mass flow prediction, a precise as possible area flow prediction is needed. Therefore, research for enhanced ML modelling should be conducted. First, it could be explored how low magnitude area flow materials can be predicted better. Here, ML model building with a focus on R^2 or a multiple indicator optimization is expected to deliver improved results. Furthermore, application of different hyperparameter tuning methodologies is recommended for research to leverage

the full potential of the applied ML models. Approaches like random search, a combination of random search and grid search or the application of more elaborated hyperparameter tuning methods resemble potential research directions.

To unlock the potentials of the developed modelling building blocks, the generalizability of the approach to the rest of the plant and to other plants need to be tested. Furthermore, the interplay of belt weigher NIR-scanner pairs for separation efficiency determination, at different places in the plant, should be researched. Together with the prediction of area flows, real-time machine optimization and simulation of impacts on the plant, through changed material input and machine settings, can be unlocked. To achieve this goal, several research areas have to be resolved. First, it is of interest how sufficient temporal resolution, with high enough accuracy for real-time plant optimization, can be obtained. In the best case, the developed approach in this work can deliver this information. If it fails to do so, alternative approaches have to be researched. Furthermore, the validity of the derived models into the future and the frequency of needed updates is of interest. When sufficient temporal resolution and accuracy is proven, the best approaches to put the sought-after real-time machine optimization into practice should be explored. Here, not only product purity, but also research regarding energy saving is recommended, to unlock greatest economic and environmental potentials.

Bibliography

Alonso, Á., Torres, A., & Dorronsoro, J. R. (2015). Random forests and gradient boosting for wind energy prediction. In Hybrid Artificial Intelligent Systems: 10th International Conference, HAIS 2015, Bilbao, Spain, June 22-24, 2015, Proceedings 10 (pp. 26-37). Springer International Publishing.

Anghel, A., Papandreou, N., Parnell, T., De Palma, A., & Pozidis, H. (2018). Benchmarking and optimization of gradient boosting decision tree algorithms. arXiv preprint arXiv:1809.04559.

Anggoro, D. A., & Mukti, S. S. (2021). Performance Comparison of Grid Search and Random Search Methods for Hyperparameter Tuning in Extreme Gradient Boosting Algorithm to Predict Chronic Kidney Failure. International Journal of Intelligent Engineering & Systems, 14(6).

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, *24*.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(2).

Berrar, D. (2019). Cross-validation.

Chen, X. (2022). Machine learning approach for a circular economy with waste recycling in smart cities. Energy Reports, 8, 3127-3140.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). Choi, B. S. (2016). An overview of bootstrapping method applicable to survey researches in rehabilitation science. Physical Therapy Korea, 23(2), 93-99.

Civancik-Uslu, D., Puig, R., Voigt, S., Walter, D., & Fullana-i-Palmer, P. (2019). Improving the production chain with LCA and eco-design: application to cosmetic packaging. *Resources, Conservation and Recycling*, *151*, 104475.

Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., ... & Cohen, I. (2009). Pearson correlation coefficient. Noise reduction in speech processing, 1-4.

Curtis, A., Küppers, B., Möllnitz, S., Khodier, K., & Sarc, R. (2021). Real-time material flow monitoring in mechanical waste processing and the relevance of fluctuations. Waste management, 120, 687-697.

Datta, P., Das, P., & Kumar, A. (2022). Hyper parameter tuning based gradient boosting algorithm for detection of diabetic retinopathy: an analytical review. Bulletin of Electrical Engineering and Informatics, 11(2), 814-824.

De Souza, S. V., & Junqueira, R. G. (2005). A procedure to assess linearity by ordinary least squares method. Analytica Chimica Acta, 552(1-2), 25-35.

Dempster, A. P., Schatzoff, M., & Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. Journal of the American Statistical Association, 72(357), 77-91.

Desai, Y., Dalvi, A., Jadhav, P., & Baphna, A. (2018). Waste segregation using machine learning. International Journal for Research in Applied Science and Engineering Technology, 6, 537-541.

Dismuke, C., & Lindrooth, R. (2006). Ordinary least squares. Methods and designs for outcomes research, 93(1), 93-104.

Dokl, M., Van Fan, Y., Vujanović, A., Pintarič, Z. N., Aviso, K. B., Tan, R. R., ... & Čuček, L. (2024). A waste separation system based on sensor technology and deep learning: A simple approach applied to a case study of plastic packaging waste. Journal of Cleaner Production, 450, 141762.

Du, W., Zheng, J., Li, W., Liu, Z., Wang, H., & Han, X. (2022). Efficient recognition and automatic sorting technology of waste textiles based on online near infrared spectroscopy and convolutional neural network. Resources, Conservation and Recycling, 180, 106157.

El-Hassani, F. Z., Amri, M., Joudar, N. E., & Haddouch, K. (2024). A new optimization model for MLP hyperparameter tuning: modelling and resolution by real-coded genetic algorithm. Neural Processing Letters, 56(2), 105.

European commission. (2018). *DIRECTIVE (EU) 2018/852 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 30 May 2018 amending Directive 2008/98/EC on waste (Text with EEA relevance).* Retrieved April 23, 2024, from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018L0852

Eurostat. (2024a). Recycling rate of municipal waste [Data files]. Retrieved April 24, 2024, from https://ec.europa.eu/eurostat/databrowser/view/cei_wm011/default/table?lang=en&category=t_env.t_env_was.t_env_wasst

Eurostat. (2024b). Recycling rates for packaging waste [Data files]. Retrieved April 24, 2024, from https://ec.europa.eu/eurostat/databrowser/view/ten00063/default/table?lang=en&category=t_env.t_env_was.t_env_wasst

Eriksen, M. K., & Astrup, T. F. (2019). Characterisation of source-separated, rigid plastic waste and evaluation of recycling initiatives: Effects of product design and source-separation system. Waste Management, 87, 161-172.

Hagquist, C., & Stenbeck, M. (1998). Goodness of fit in regression analysis–R 2 and G 2 reconsidered. Quality and Quantity, 32(3), 229-245.

Hoque, K. E., & Aljamaan, H. (2021). Impact of hyperparameter tuning on machine learning models in stock price forecasting. IEEE Access, 9, 163815-163830.

Fan, M., Xiao, K., Sun, L., Zhang, S., & Xu, Y. (2022). Automated Hyperparameter Optimization of Gradient Boosting Decision Tree Approach for Gold Mineral Prospectivity Mapping in the Xiong'ershan Area. Minerals, 12(12), 1621.

Feil, A., van Velzen, E. T., Jansen, M., Vitz, P., Go, N., & Pretz, T. (2016). Technical assessment of processing plants as exemplified by the sorting of beverage cartons from lightweight packaging wastes. *Waste management*, *48*, 95-105.

Feil, A., Pretz, T., Vitz, P., & Thoden van Velzen, E. U. (2017). A methodical approach for the assessment of waste sorting plants. Waste Management & Research, 35(2), 147-154.

Friedrich, K., Koinig, G., Pomberger, R., & Vollprecht, D. (2022). Qualitative analysis of post-consumer and post-industrial waste via near-infrared, visual and induction identification with experimental sensor-based sorting setup. MethodsX, 9, 101686.

Gadaleta, G., De Gisi, S., Binetti, S. M., & Notarnicola, M. (2020). Outlining a comprehensive techno-economic approach to evaluate the performance of an advanced sorting plant for plastic waste recovery. Process Safety and Environmental Protection, 143, 248-261.

Gujarati, D. N. (2021). Multicollinearity. In Basic econometrics (5th ed., pp. [343-375]). McGraw-Hill.
Itano, F., de Sousa, M. A. D. A., & Del-Moral-Hernandez, E. (2018, July). Extending MLP ANN hyper-parameters Optimization by using Genetic Algorithm. In 2018 International joint conference on neural networks (IJCNN) (pp. 1-8). IEEE.

Kavzoglu, T., & Teke, A. (2022). Advanced hyperparameter optimization for improved spatial prediction of shallow landslides using extreme gradient boosting (XGBoost). Bulletin of Engineering Geology and the Environment, 81(5), 201.

Kaza, S., Yao, L., Bhada-Tata, P., & Van Woerden, F. (2018). *What a waste 2.0: a global snapshot of solid waste management to 2050*. World Bank Publications.

Kroell, N., Chen, X., Maghmoumi, A., Koenig, M., Feil, A., & Greiff, K. (2021). Sensor-based particle mass prediction of lightweight packaging waste using machine learning algorithms. *Waste management*, *136*, 253-265.

Kroell, N., Chen, X., Greiff, K., & Feil, A. (2022a). Optical sensors and machine learning algorithms in sensor-based material flow characterization for mechanical recycling processes: A systematic literature review. Waste Management, 149, 259-290.

Kroell, N., Maghmoumi, A., Dietl, T., Chen, X., Küppers, B., Scherling, T., ... & Greiff, K. (2024a). Towards digital twins of waste sorting plants: Developing data-driven process models of industrial-scale sensor-based sorting units by combining machine learning with near-infrared-based process monitoring. Resources, Conservation and Recycling, 200, 107257.

Kroell, N., Chen, X., Küppers, B., Schlögl, S., Feil, A., & Greiff, K. (2024b). Near-infrared-based quality control of plastic pre-concentrates in lightweight-packaging waste sorting plants. Resources, Conservation and Recycling, 201, 107256.

Kuhn, M., & Johnson, K. (2013). Over-fitting and model tuning. Applied predictive modelling, 61-92.
Küppers, B., Seidler, I., Koinig, G., Pomberger, R., & Vollprecht, D. (2020). Influence of throughput rate and input composition on sensor-based sorting efficiency. Detritus, 9(March), 59-67.

La Tour, T. D., Eickenberg, M., Nunez-Elizalde, A. O., & Gallant, J. L. (2022). Feature-space selection with banded ridge regression. NeuroImage, 264, 119728.

Maga, D., Hiebel, M., & Aryan, V. (2019). A comparative life cycle assessment of meat trays made of various packaging materials. *Sustainability*, *11*(19), 5324.

Maier, G., Pfaff, F., Pieper, C., Gruna, R., Noack, B., Kruggel-Emden, H., ... & Beyerer, J. (2020). Experimental evaluation of a novel sensor-based sorting approach featuring predictive real-time multiobject tracking. *IEEE Transactions on Industrial Electronics*, *68*(2), 1548-1559.

Mokhtar, S. F., Yusof, Z. M., & Sapiri, H. (2023). Confidence intervals by bootstrapping approach: a significance review. Malaysian Journal of Fundamental and Applied Sciences, 19(1), 30-42.

Moksony, F., & Heged, R. (1990). Small is beautiful. The use and interpretation of R2 in social research. Szociológiai Szemle, Special issue, 130-138.

Naidu, G., Zuva, T., & Sibanda, E. M. (2023, April). A review of evaluation metrics in machine learning algorithms. In Computer Science On-line Conference (pp. 15-25). Cham: Springer International Publishing.

Nugroho, W. H., Handoyo, S., Hsieh, H. C., Akri, Y. J., & DwinitaAdelia, D. (2022). Modelling multioutput response uses ridge regression and MLP neural network with tuning hyperparameter through cross validation. International Journal of Advanced Computer Science and Applications, 13(9).

Pan, S., Zheng, Z., Guo, Z., & Luo, H. (2022). An optimized XGBoost method for predicting reservoir porosity using petrophysical logs. Journal of Petroleum Science and Engineering, 208, 109520.

Ogunsanya, M., Isichei, J., & Desai, S. (2023). Grid search hyperparameter tuning in additive manufacturing processes. Manufacturing Letters, 35, 1031-1042.

Ørebæk, O. E., & Geitle, M. (2021). Exploring the Hyperparameters of XGBoost Through 3D Visualizations. In AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering.

Ozdemir, E., M., Ali, Z., Subeshan, B., & Asmatulu, E. (2021). Applying machine learning approach in recycling. Journal of Material Cycles and Waste Management, 23, 855-871.

Paul, R. K. (2006). Multicollinearity: Causes, effects and remedies. IASRI, New Delhi, 1(1), 58-65.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Pekel, E. (2020). Estimation of soil moisture using decision tree regression. Theoretical and Applied Climatology, 139(3), 1111-1119.

Probst, P., Boulesteix, A. L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. Journal of Machine Learning Research, 20(53), 1-32.

Rahman, M. W., Islam, R., Hasan, A., Bithi, N. I., Hasan, M. M., & Rahman, M. M. (2022). Intelligent waste management system using deep learning with IoT. Journal of King Saud University-Computer and Information Sciences, 34(5), 2072-2087.

Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. Scientific Reports, 14(1), 6086.

Rokem, A., & Kay, K. (2020). Fractional ridge regression: a fast, interpretable reparameterization of ridge regression. GigaScience, 9(12), giaa133.

Sahin, E. K. (2020). Assessing the predictive capability of ensemble tree methods for landslide susceptibility mapping using XGBoost, gradient boosting machine, and random forest. SN Applied Sciences, 2(7), 1308.

Schwarz, A. E., Ligthart, T. N., Bizarro, D. G., De Wild, P., Vreugdenhil, B., & Van Harmelen, T. (2021). Plastic recycling in a circular economy; determining environmental performance through an LCA matrix model approach. *Waste Management*, *121*, 331-342.

Sheng, T. J., Islam, M. S., Misran, N., Baharuddin, M. H., Arshad, H., Islam, M. R., ... & Islam, M. T. (2020). An internet of things based smart waste management system using LoRa and tensorflow deep learning model. IEEE Access, 8, 148793-148811.

Shrestha, N. (2020). Detecting multicollinearity in regression analysis. American Journal of Applied Mathematics and Statistics, 8(2), 39-42.

Sutco. (n.d.). SORTIER- UND AUFBEREITUNGSANLAGEN FÜR DIE WELTWEITE RECYCLINGWIRTSCHAFT. Retrieved October 10, 2024, from https://www.sutco.com/de/.

Sutco. (2024a). Aktuelles – Neuste Meldungen: Inbetriebnahme der ersten Sortieranlage für Abfälle in Chile; "Ennshafen - Sutco is setting new standards in the recycling industry"; Modern waste sorting plant for ProNatura Bydgoszcz. Retrieved September 30, 2024, from https://www.sutco.com/de/aktuelles.

Sutco. (2024b). ProDIGIT – Digitale Lösungen für Sortieranlagen. Retrieved September 30, 2024, from https://www.sutco.com/de/produkte/trockenmechanik/prodigit.

Tanguay-Rioux, F., Legros, R., & Spreutels, L. (2021). On the limits of empirical partition coefficients for modelling material recovery facility unit operations in municipal solid waste management. Journal of cleaner production, 293, 126016.

Tanguay-Rioux, F., Provost-Savard, A., Spreutels, L., Héroux, M., & Legros, R. (2022). A method for assessing the performance of sorting unit operations in a material recovery facility based on waste characterizations. *The Canadian Journal of Chemical Engineering*, *100*(9), 2572-2586.

Tarwidi, D., Pudjaprasetya, S. R., Adytia, D., & Apri, M. (2023). An optimized XGBoost-based machine learning method for predicting wave run-up on a sloping beach. MethodsX, 10, 102119.

TOMRA. (2024). AA104 – PO75+; AA108 – PO75 Clean. Retrieved July 15, 2024, from https://bp.insight.tomra.com/machine-details

Tonini, D., Schrijvers, D., Nessi, S., Garcia-Gutierrez, P., & Giuntoli, J. (2021). Carbon footprint of plastic from biomass and recycled feedstock: methodological insights. *The International Journal of Life Cycle Assessment*, *26*, 221-237.

TU Dresden. (2023). EnSort - "Steigerung der Energieeffizi-enz im Abfall- und Recyclingstoff-Sortierprozess durch Erhöhung des Automatisie-rungsgrades". Retrieved September 30, 2024, from https://tu-dresden.de/ing/elektrotechnik/ifa/plt/forschung/forschungsprojekte/ensort-steigerung-der-energieeffizienz-im-abfall-und-recyclingstoff-sortierprozess-durch-erhoehung-des-automatisierungsgrades.

Vrancken, C., Longhurst, P. J., & Wagland, S. T. (2017). Critical review of real-time methods for solid waste characterisation: Informing material recovery and fuel production. Waste management, 61, 40-57.

Waste Management World (WMW). (2024). Sutco is setting new standards in the recycling industry. Retrieved October 23, 2024, from https://waste-management-world.com/resource-use/sutco-is-setting-new-standards-in-the-recycling-industry/.

Xi, H., Li, Z., Han, J., Shen, D., Li, N., Long, Y., ... & Liu, H. (2022). Evaluating the capability of municipal solid waste separation in China based on AHP-EWM and BP neural network. Waste Management, 139, 208-216.

Xu, M., Watanachaturaporn, P., Varshney, P. K., & Arora, M. K. (2005). Decision tree regression for soft classification of remote sensing data. Remote Sensing of Environment, 97(3), 322-336.

Youness, G., Phan, N. U. T., & Boulakia, B. C. (2023, December). BootBOGS: Hands-on optimizing Grid Search in hyperparameter tuning of MLP. In AICCSA 2023: 20th ACS/IEEE International Conference on Computer Systems and Applications.

Yoon, H. (2021). Finding unexpected test accuracy by cross validation in machine learning. International Journal of Computer Science & Network Security, 21(12spc), 549-555.

Young, M. T., Hinkle, J., Ramanathan, A., & Kannan, R. (2018, September). Hyperspace: Distributed bayesian hyperparameter optimization. In 2018 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD) (pp. 339-347). IEEE.

Zheng, Y., Bai, J., Xu, J., Li, X., & Zhang, Y. (2018). A discrimination model in waste plastics sorting using NIR hyperspectral imaging system. Waste Management, 72, 87-98.

Zou, M., Jiang, W. G., Qin, Q. H., Liu, Y. C., & Li, M. L. (2022). Optimized XGBoost model with small dataset for predicting relative density of Ti-6Al-4V parts manufactured by selective laser melting. Materials, 15(15), 5298.

Appendix 1: Material-specific distributions

In the following, all material-specific distributions in form of histograms can be found. The decision if values for PO75 or PO75C were made transparent was made material by material favouring plots where the least information was lost. Due to the amount of plots the figure is spread over two pages.

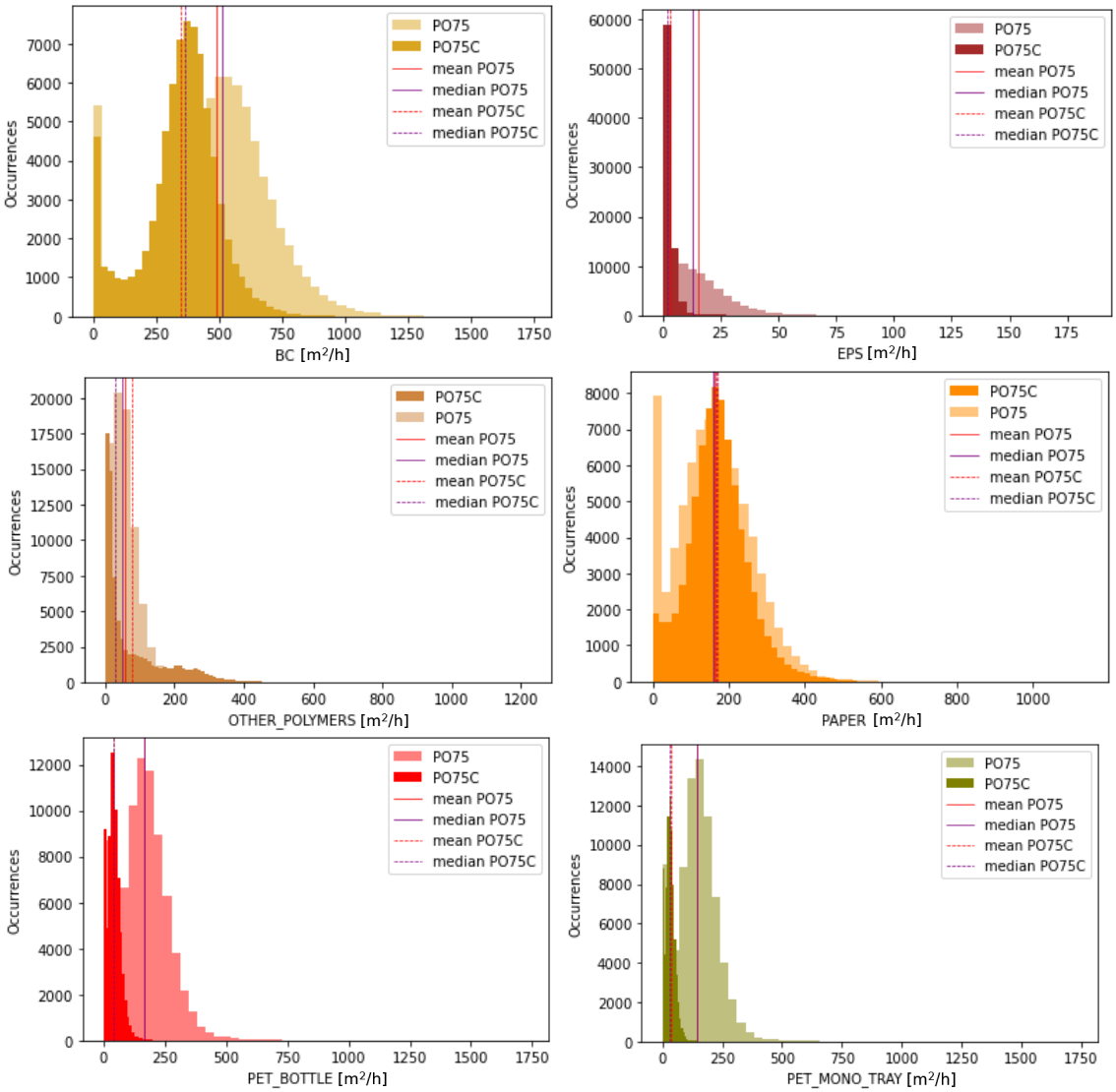

Figure A1.1: Material-specific area flow distribution for all available material, zeros have been removed upfront.

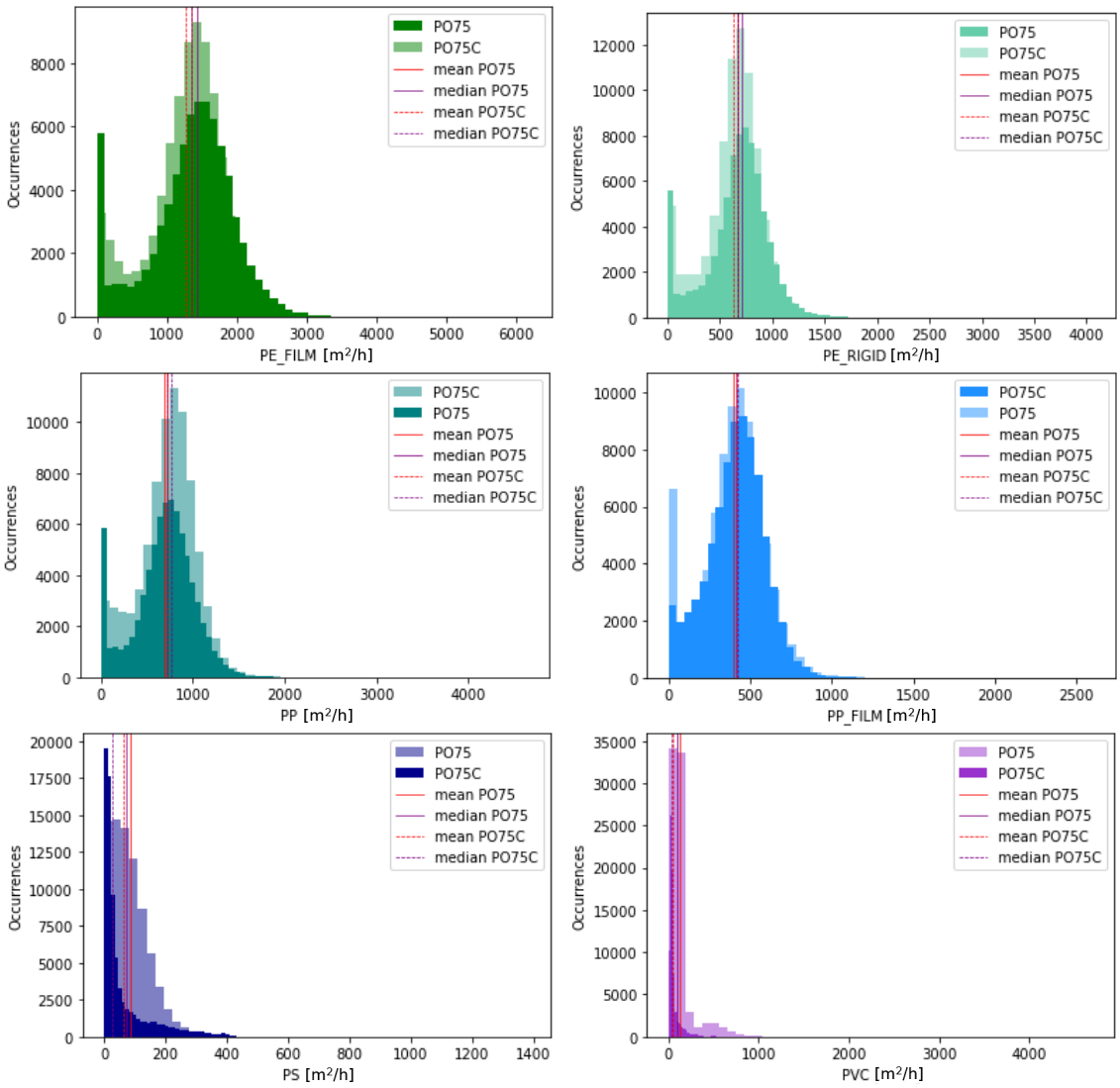

Figure A1.2: Material-specific area flow distribution for all available material, zeros have been removed upfront.

Appendix 2: Material-specific correlations

Below all material-specific correlations in form of scatter plots can be found. Due to the amount of plots the figure is spread over two pages.

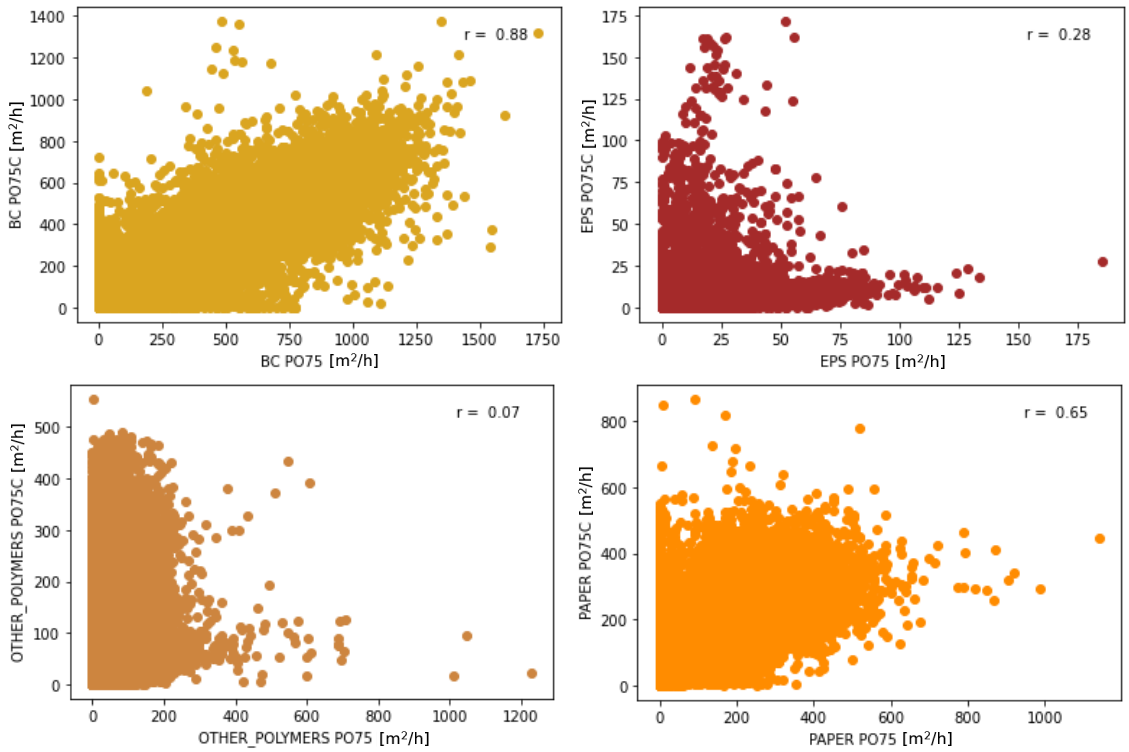

Figure A2.1: Material-specific correlations in form of scatter plots for all available material, zeros have been removed upfront.

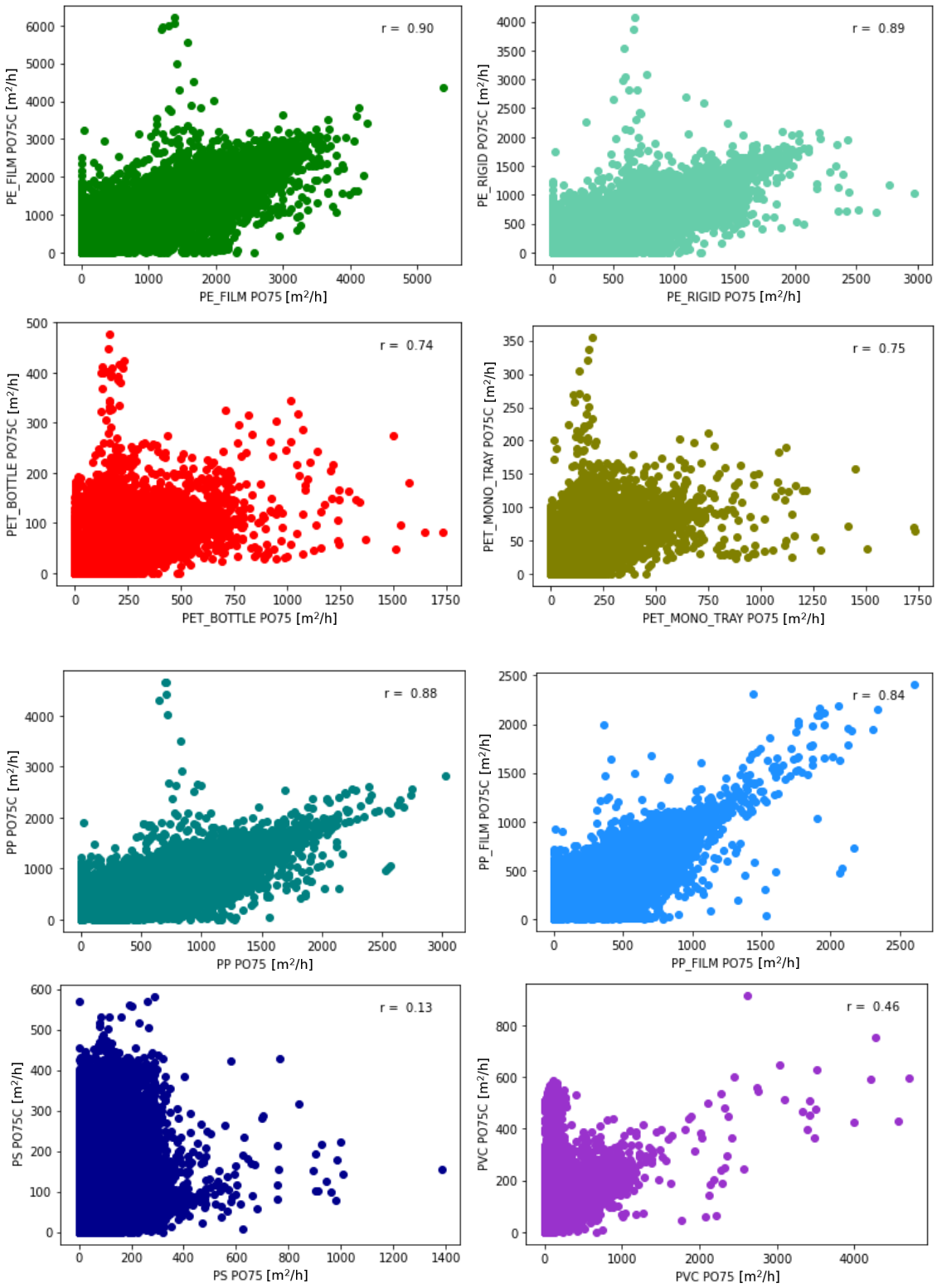

Figure A2.2: Material-specific correlations in form of scatter plots for all available material, zeros have been removed upfront.

Appendix 3: Material-specific correlation computed by time

The subsequent plots present all material-specific correlations over time. For this, the data was split up into 200 bins by time. Due to the amount of plots the figure is spread over two pages.

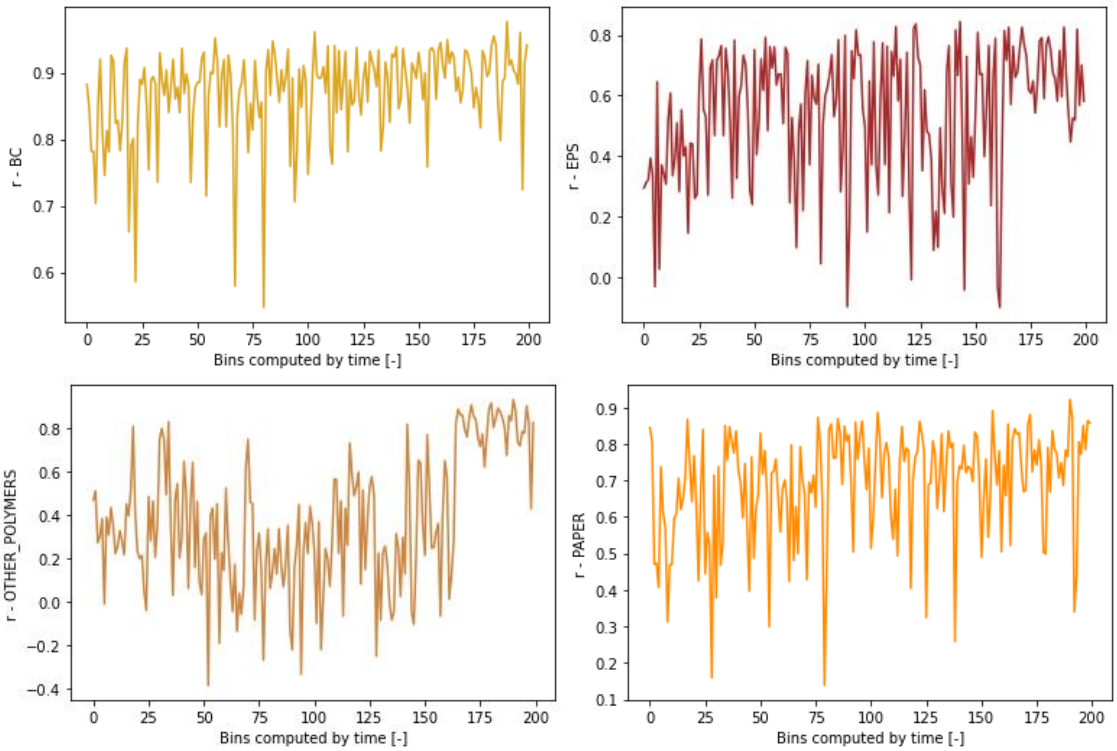

Figure A3.1: Pearson correlation coefficient for all available materials regarding NIR-scanner PO75 and PO75C with 200 bins compiled by time.

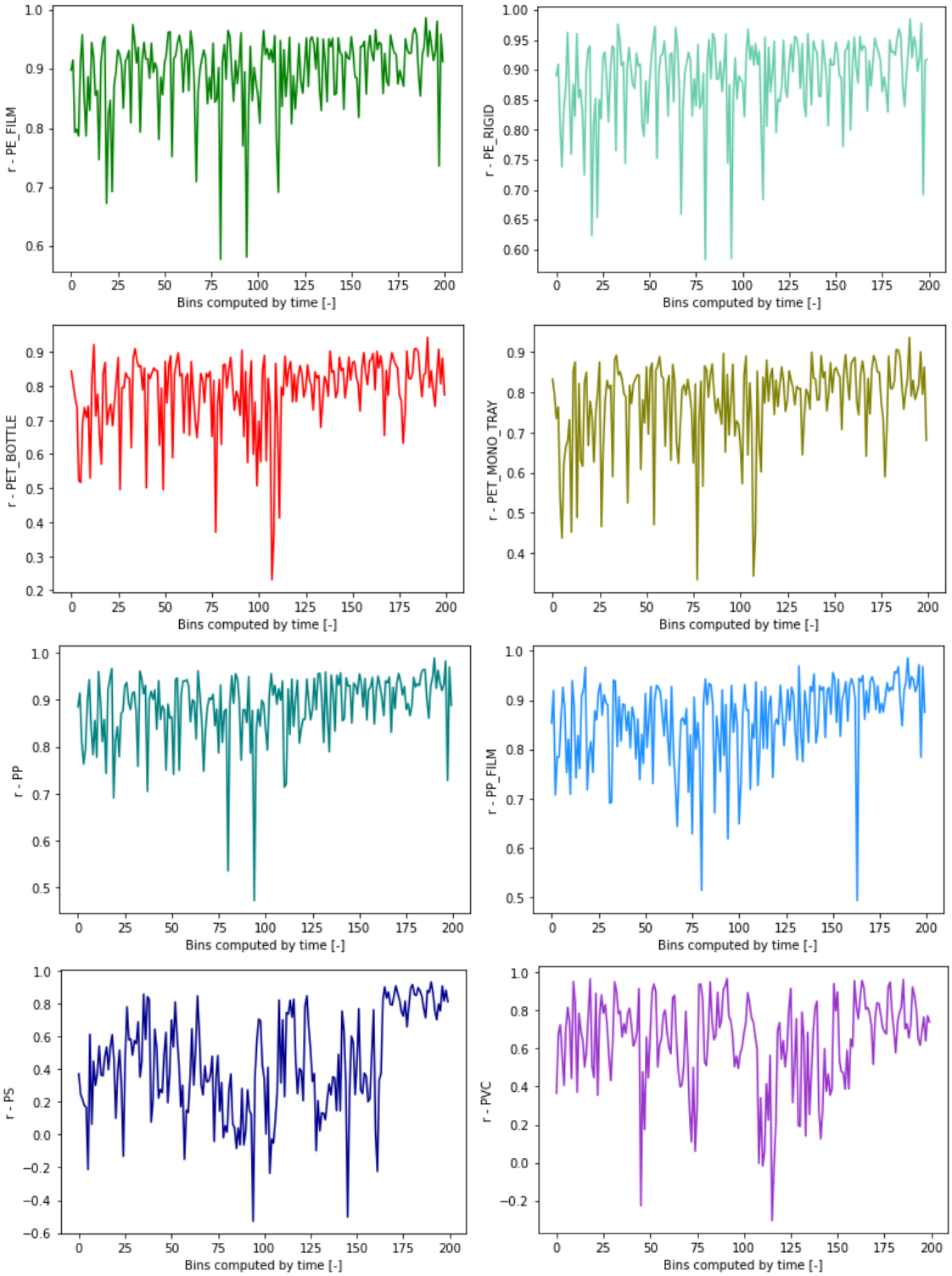

Figure A3.2: Pearson correlation coefficient for all available materials regarding NIR-scanner PO75 and PO75C with 200 bins compiled by time.

Appendix 4: Material-specific correlations of lowest performing correlation bins

In the following, all material-specific correlations, computed for the lowest correlation bin, can be found. Due to the amount of plots the figure is spread over two pages.

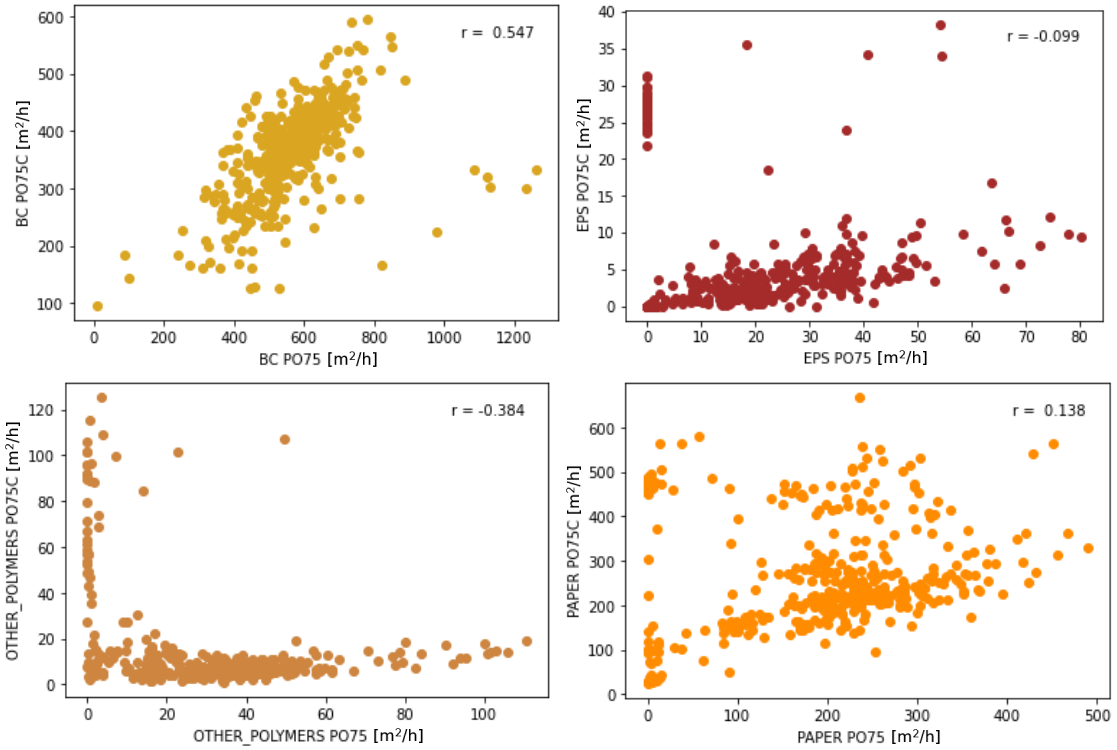

Figure A4.1: Material-specific correlations in form of scatter plots for the lowest performing correlation bin for all available material, zeros have been removed upfront.

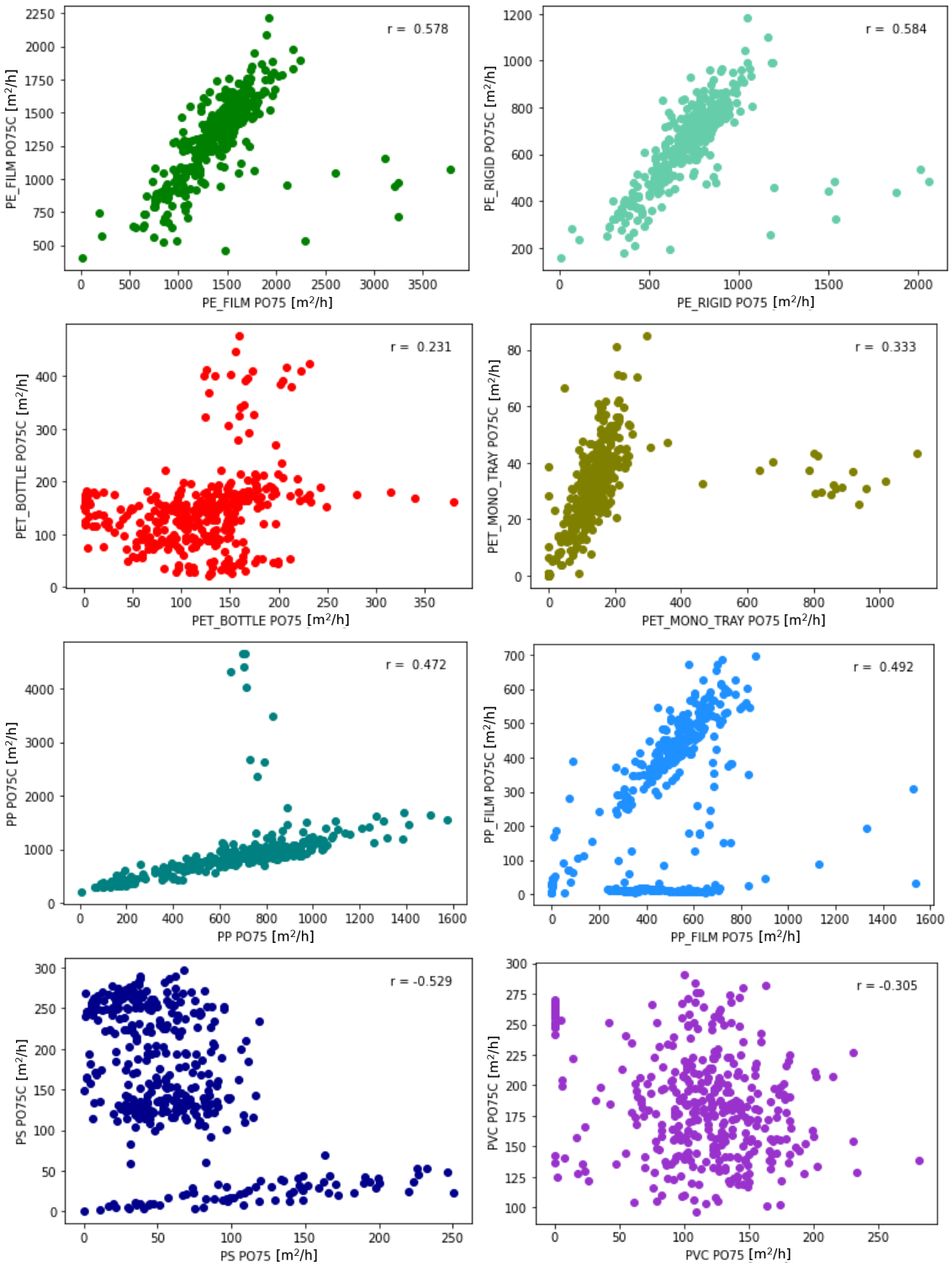

Figure A4.2: Material-specific correlations in form of scatter plots for the lowest performing correlation bin for all available material, zeros have been removed upfront.

Appendix 5: Sum PO75, Sum PO75C and material for PO75 and PO75C plot for the lowest correlation bin by time

Below, all material-specific plots for the sum of PO75 and PO75C as well as the material-specific area flows for all materials computed for the lowest correlation bin can be found. Due to the amount of plots the figure is spread over several pages.

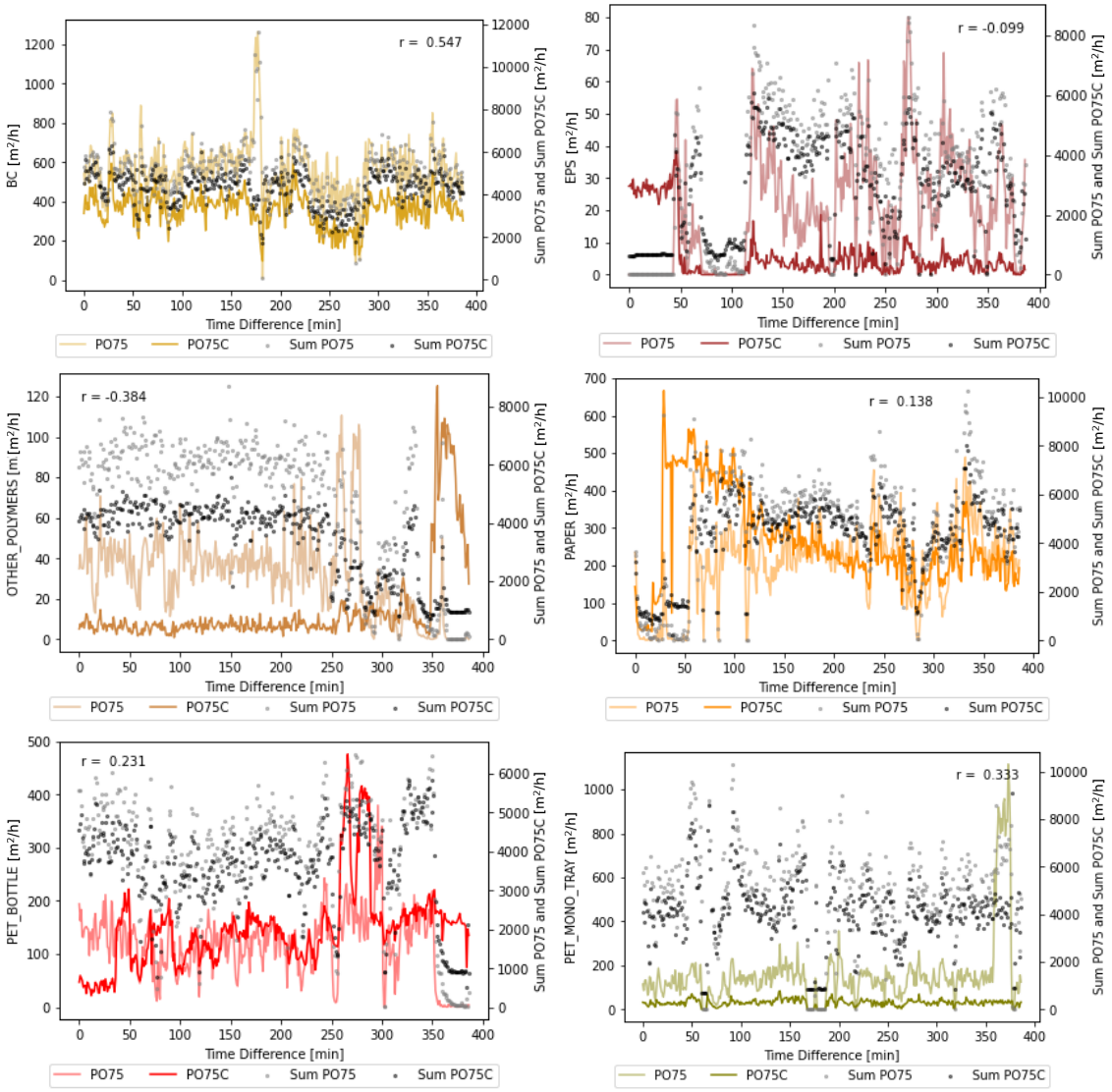

Figure A5.1: Sum of areas on PO75 (grey dots) and PO75C (black dots) as well as for the examined materials for PO75 (light lines) and PO75C (solid lines).

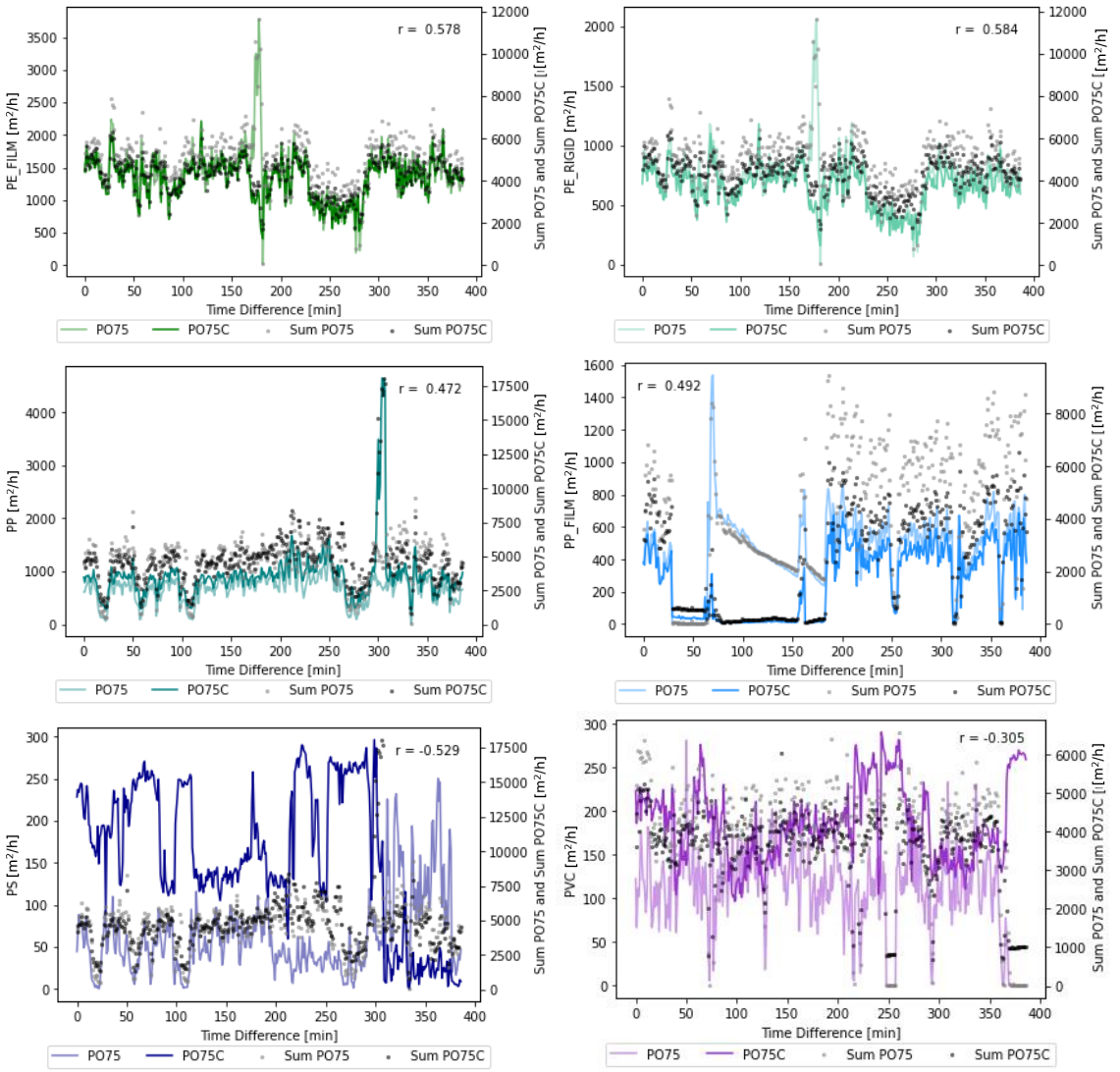

Figure A5.2: Sum of areas on PO75 (grey dots) and PO75C (black dots) as well as for the examined materials for PO75 (light lines) and PO75C (solid lines).

Appendix 6: Conveyor belt occupancies of PO75 and material occurrence for all materials for PO75 and PO75C

The subsequent plots present material-specific area flows and belt occupancies for PO75 (blue color palette) and material-specific area flows from PO75C together with belt occupancies of PO75 (red color palette). Furthermore, a full coloration and a 0.5% coloration version of the plots is provided. For better overview, the occupancies were separated in 5 categories. These categories are 0-50 [m^2/h], 50-100 [m^2/h], 100-200 [m^2/h], 200-300 [m^2/h] and >300 [m^2/h]. On the x-axis the number of occurrences per category is indicated. On the y-axis the corresponding material-specific area flow is presented. Due to the amount of plots the figure is spread over two pages.

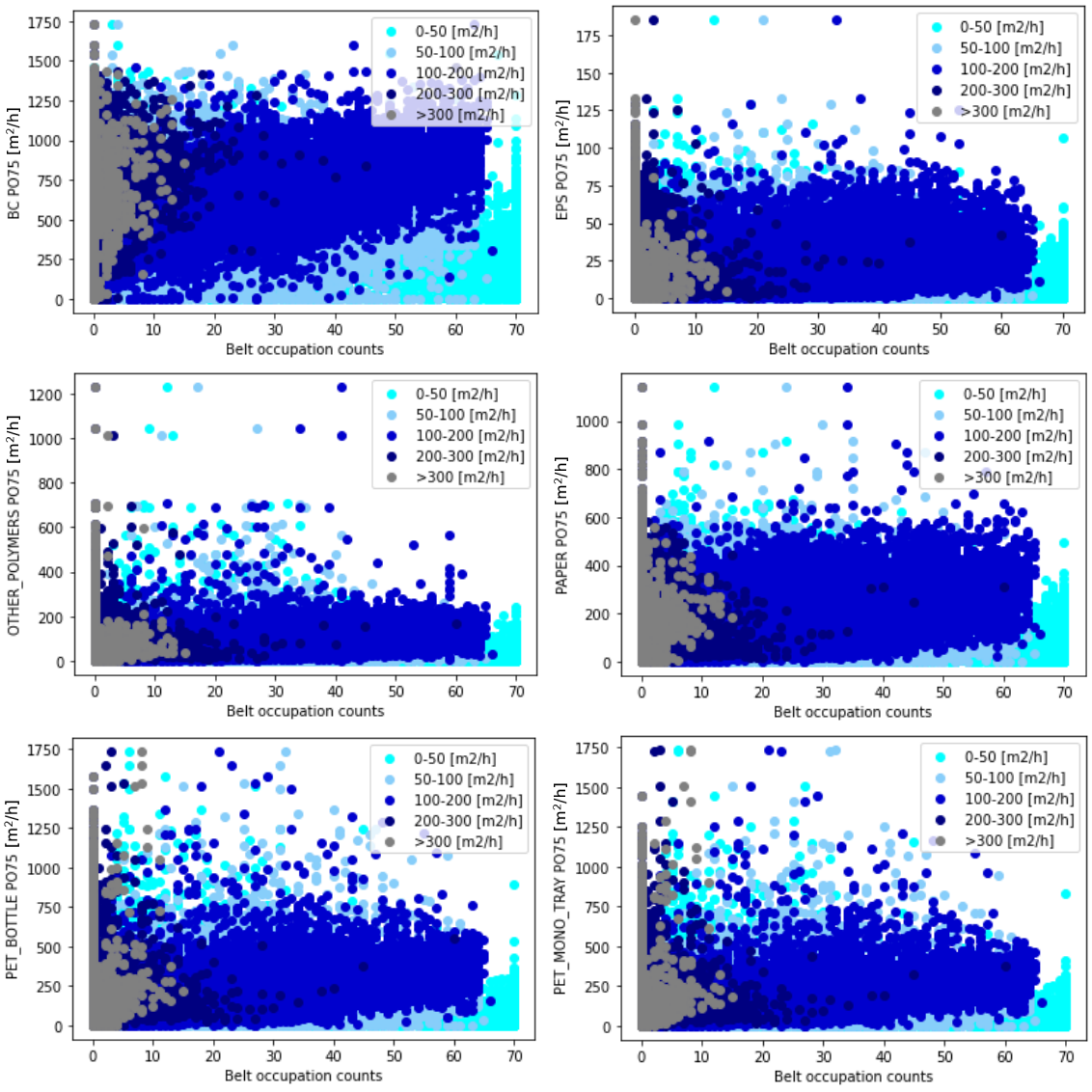

Figure A6.1: Scatter plot for all materials on PO75 together with belt occupation counts of PO75.

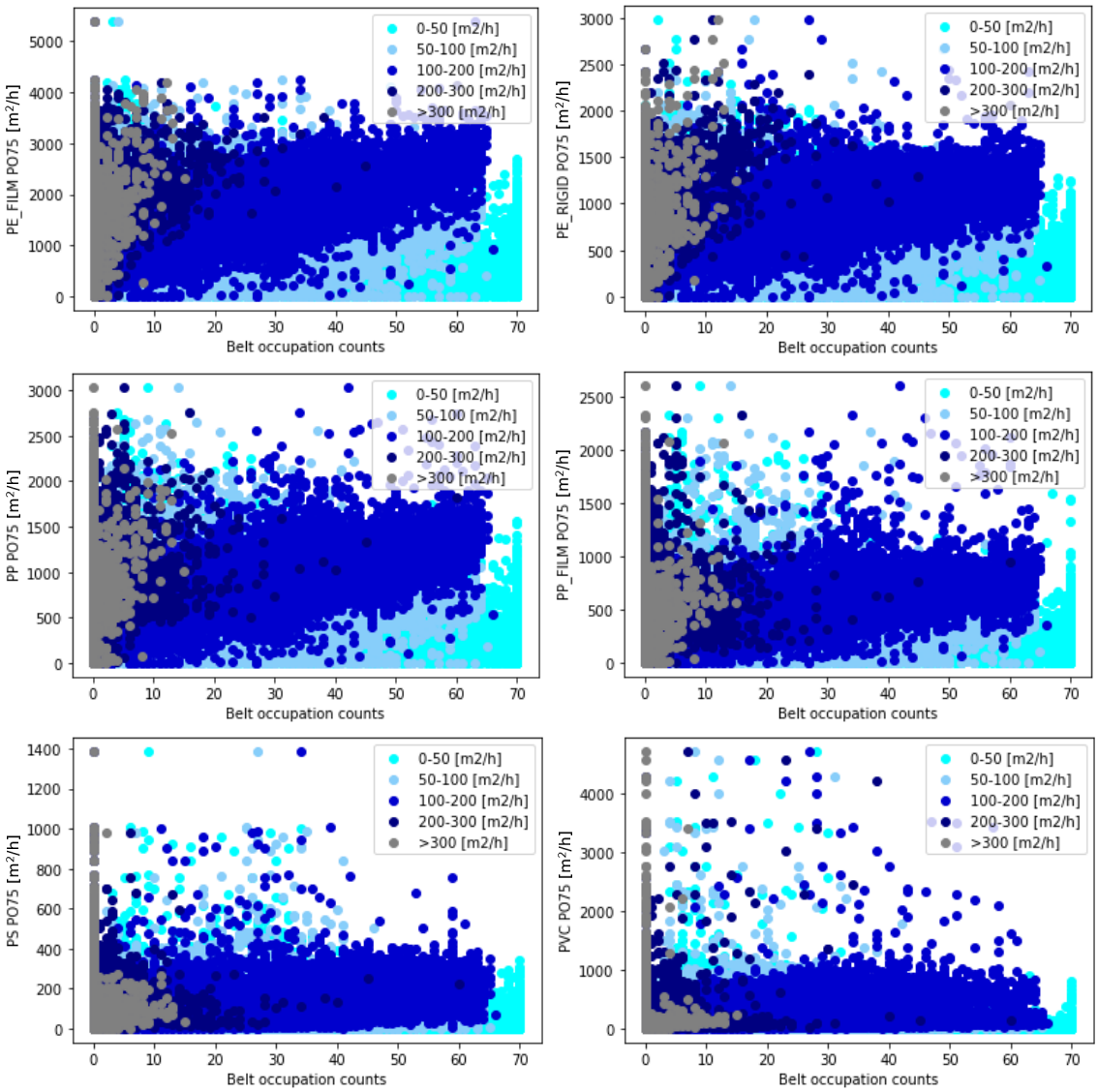

Figure A6.2: Scatter plot for all materials on PO75 together with belt occupation counts of PO75.

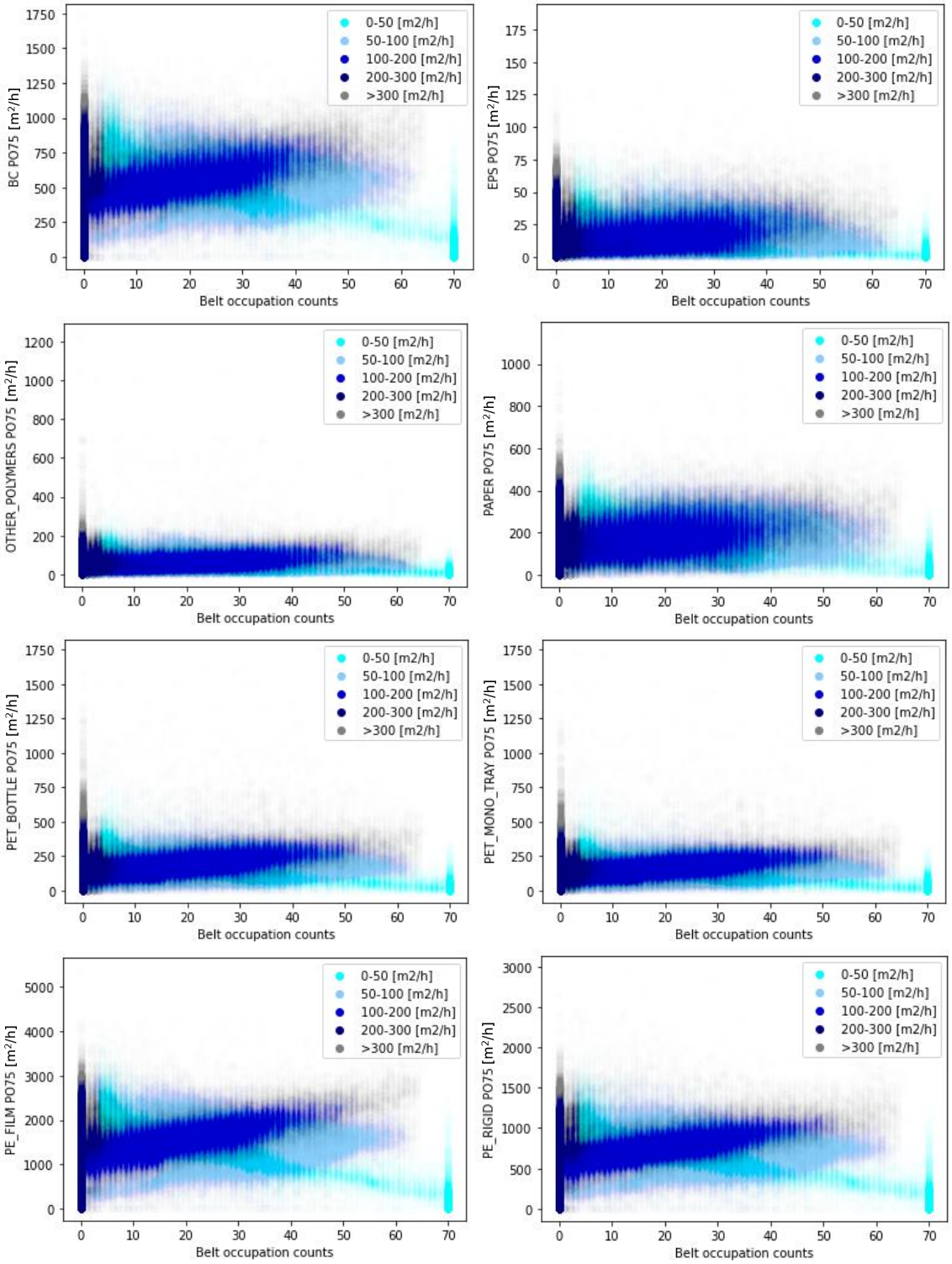

Figure A7.1: Scatter plot for all materials on PO75 together with belt occupation counts on PO75 with 0.5% coloration.

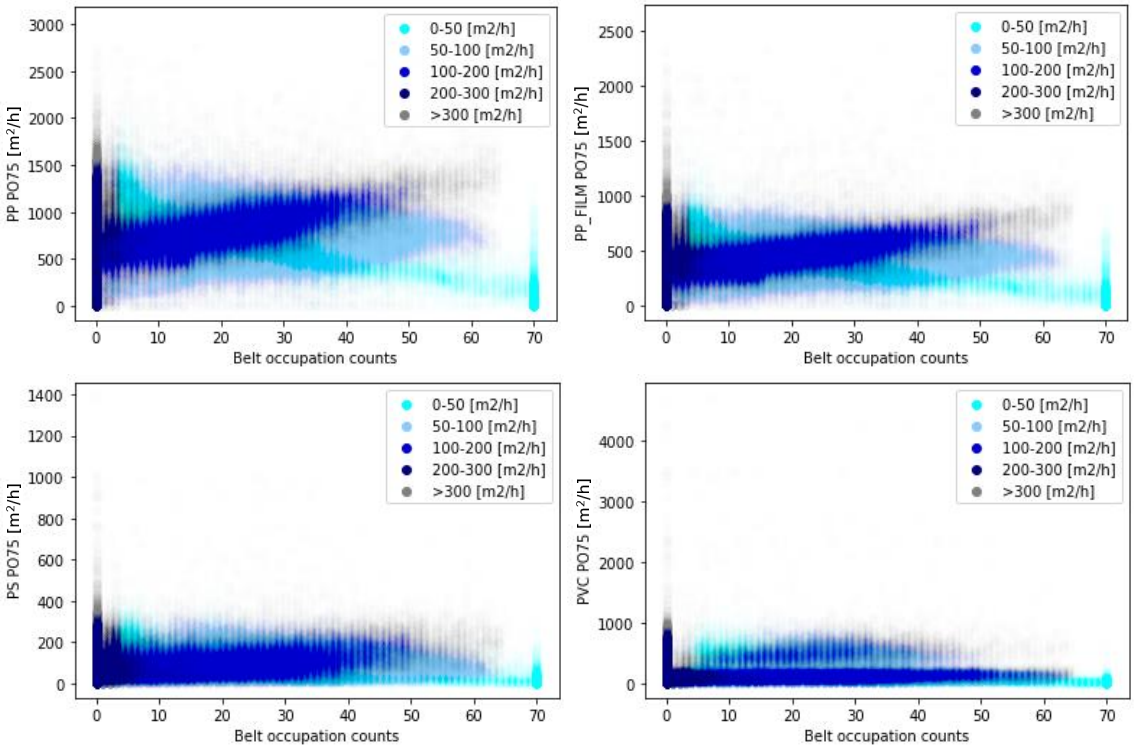

Figure A7.2: Scatter plot for all materials on PO75 together with belt occupation counts on PO75 with 0.5% coloration.

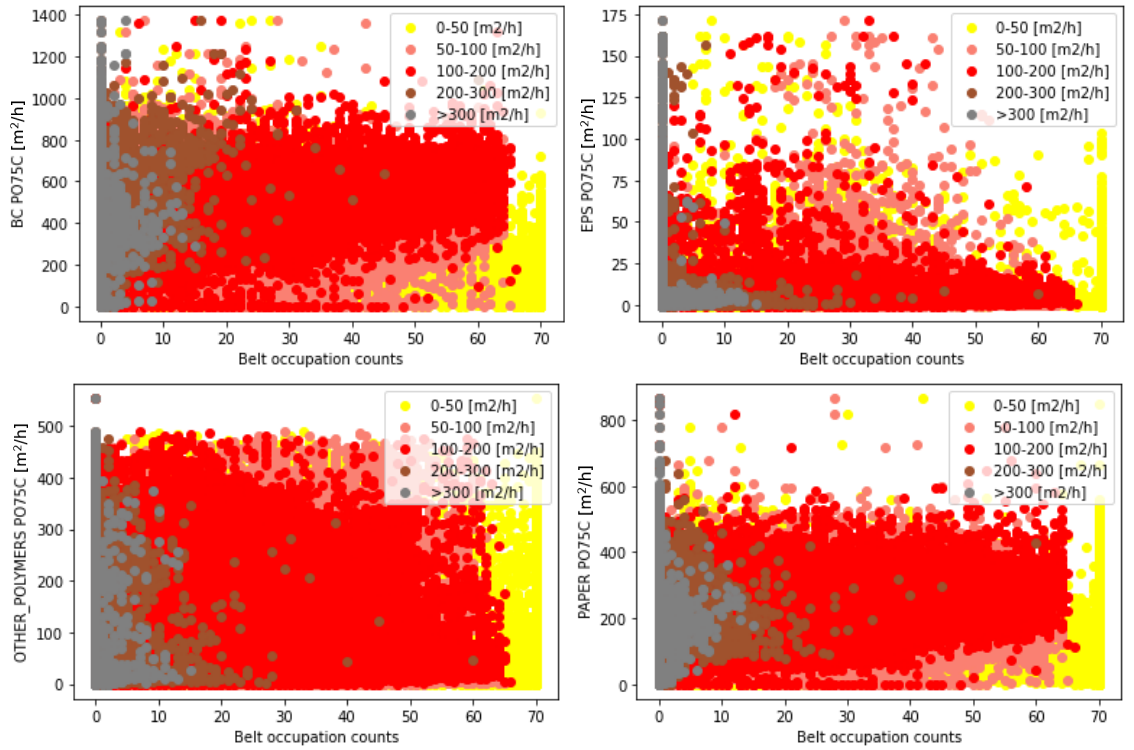

Figure A8.1: Scatter plot for all materials on PO75C together with belt occupation counts of PO75.

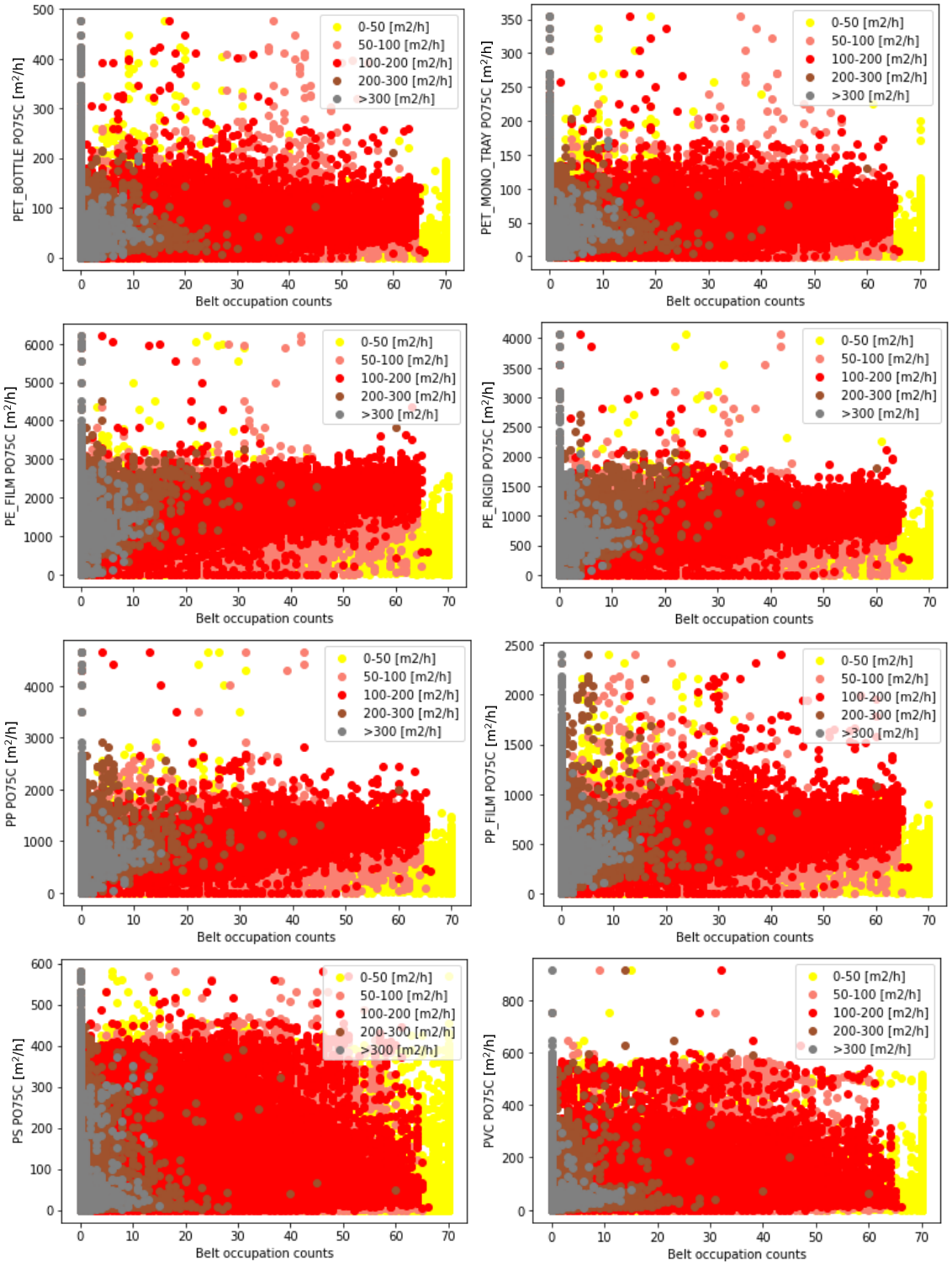

Figure A8.2: Scatter plot for all materials on PO75C together with belt occupation counts of PO75.

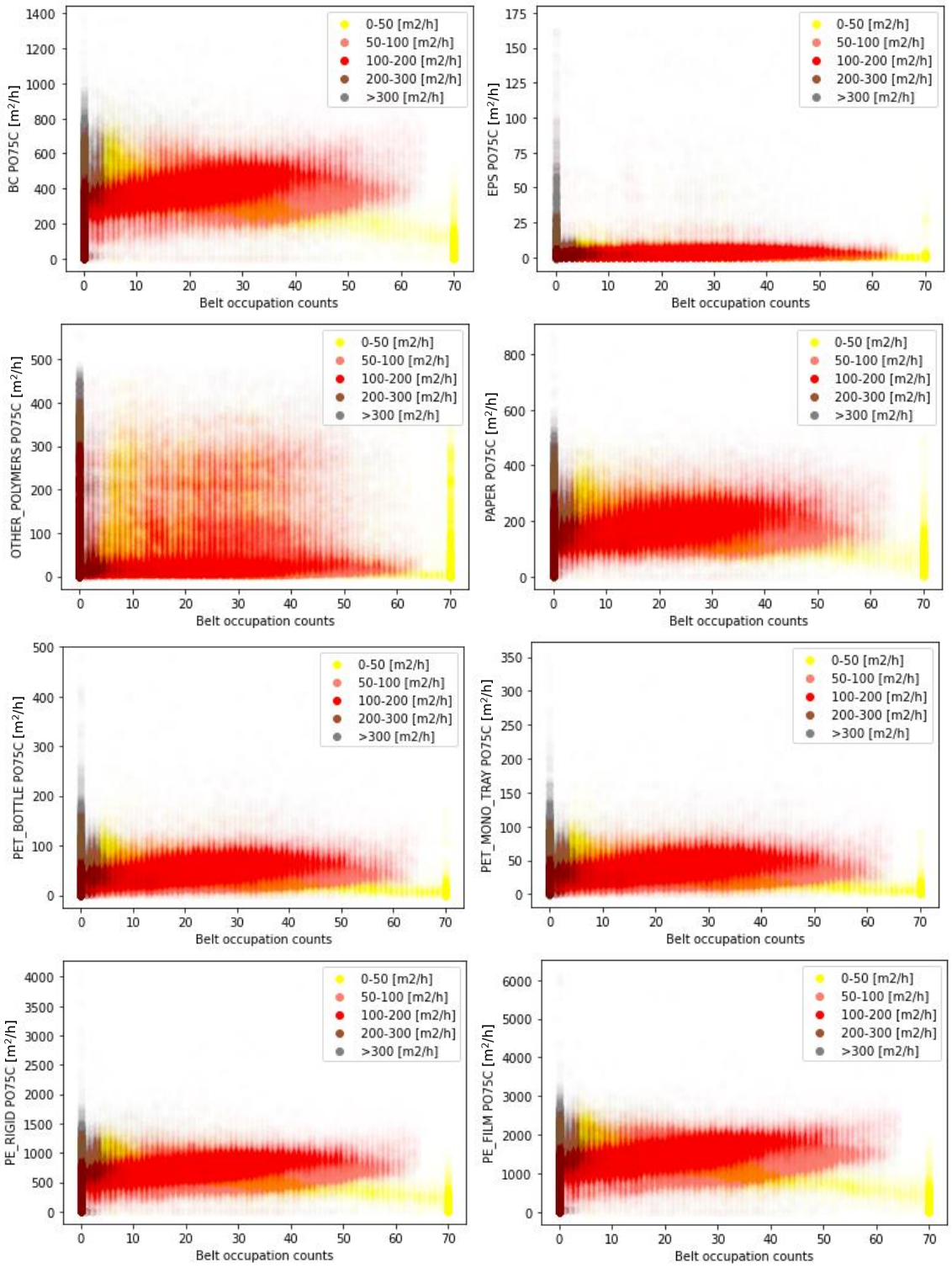

Figure A9.1: Scatter plot for all materials on PO75C together with belt occupation counts on PO75 with 0.5% coloration.

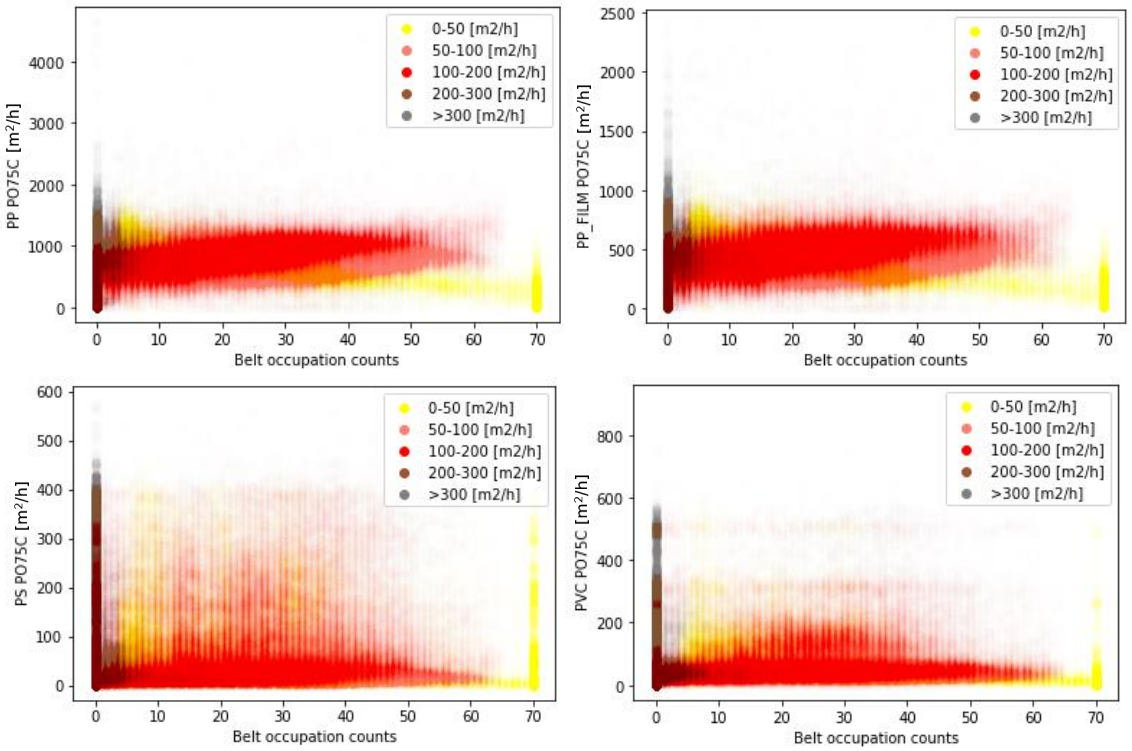

Figure A9.2: Scatter plot for all materials on PO75C together with belt occupation counts on PO75 with 0.5% coloration.

Appendix 7: Machine Learning Model Performance for V7 and V9 measured in MAE for all materials

Below, the performance of all machine learning models tested during the try-out can be found. Pre-processing versions V7 and V9 are indicated and on the y-axis the MAE can be observed. Due to the amount of plots the figure is spread over two pages.

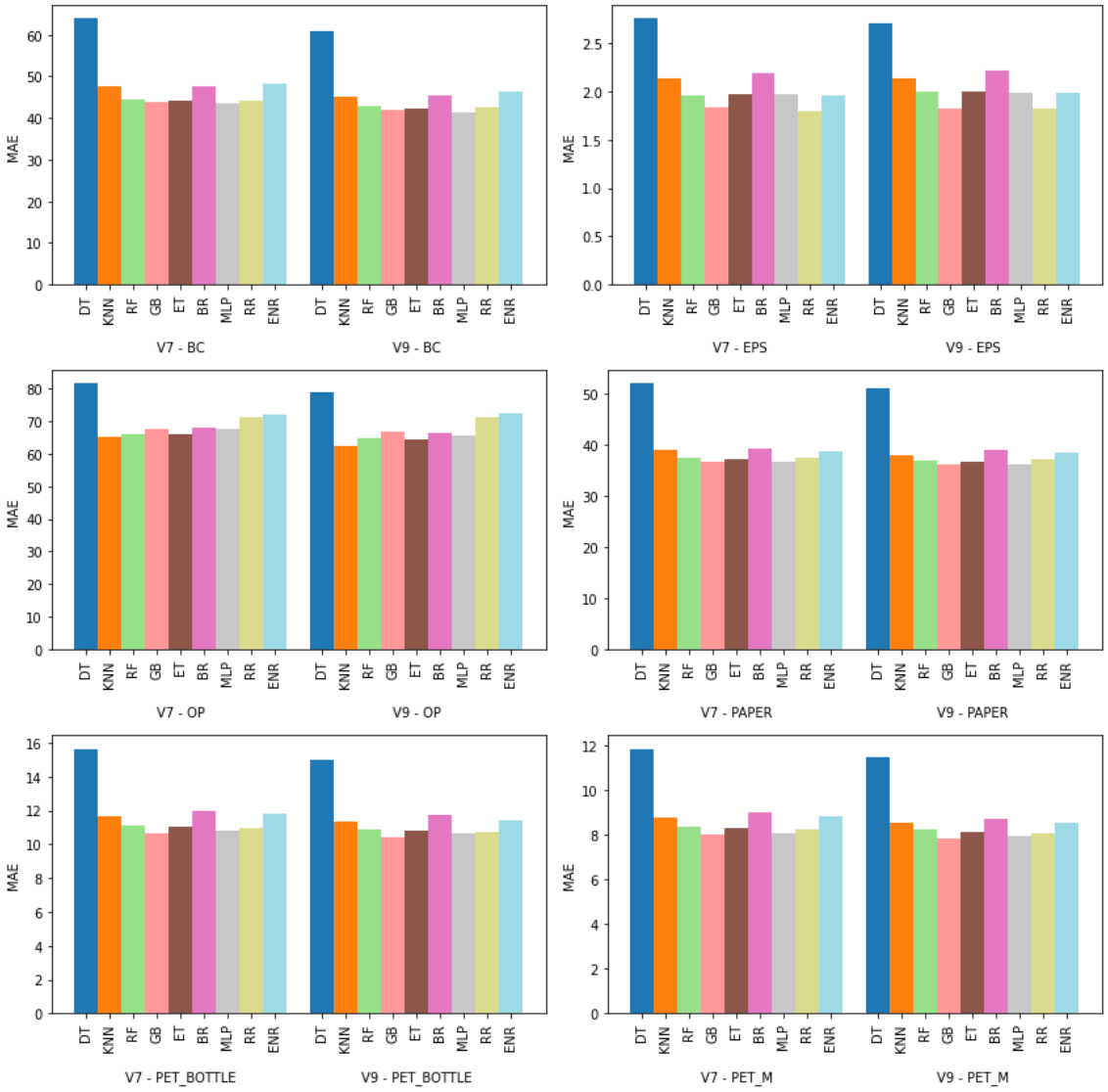

Figure A10.1: ML model try-out results for V7 and V9 data pre-processing with MAE as performance indicator.

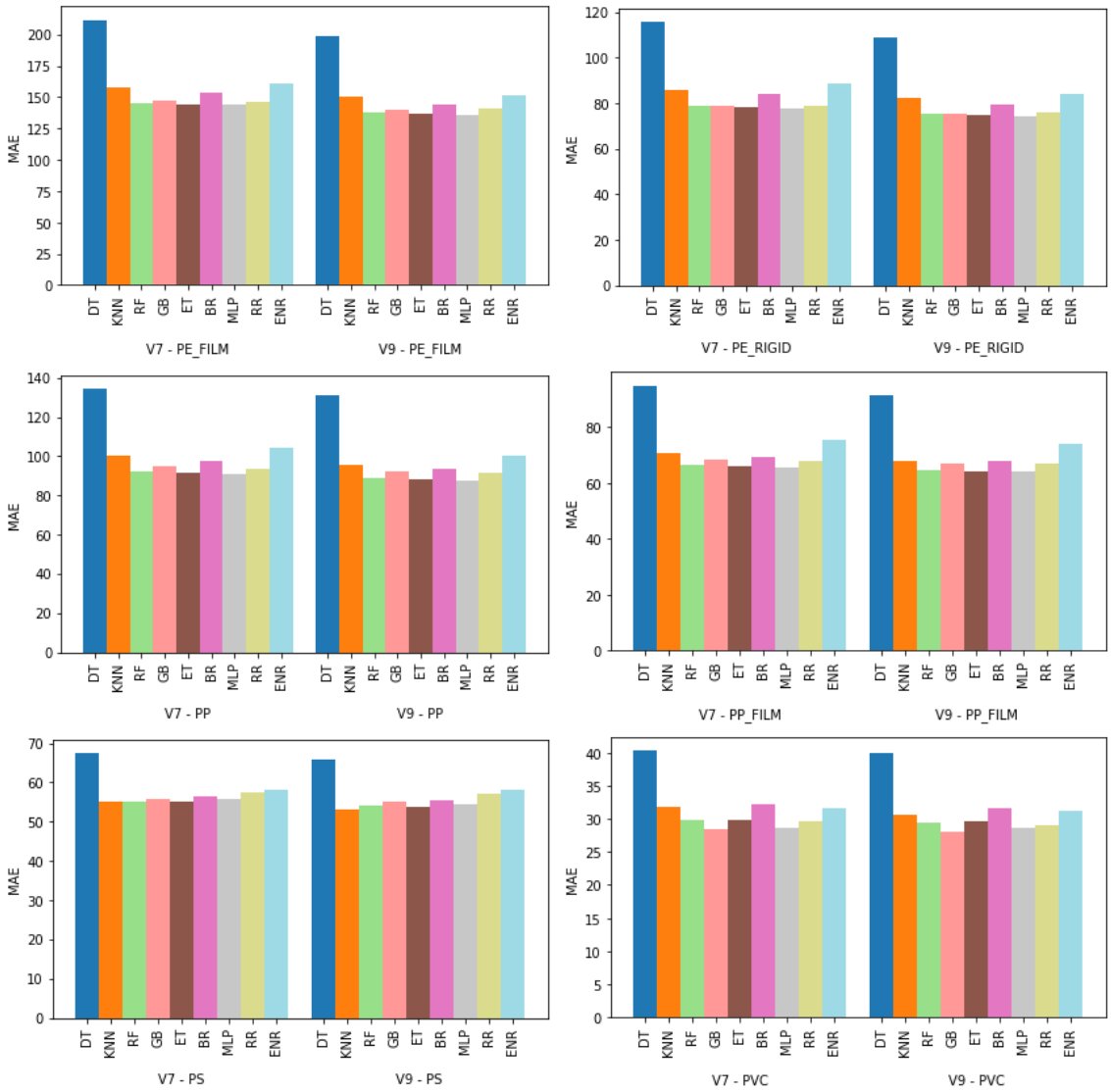

Figure A10.2: ML model try-out results for V7 and V9 data pre-processing with MAE as performance indicator.

Appendix 8: Area densities with V9 pre-processing for PO75 and AA101

Below, important figures and information for the area densities determined for PO75, with belt weigher data from AA101 and V9 data pre-processing, can be found.

Table A1: OLS results for area density prediction for PO75 using data from PO75 and AA101 after similar data pre-processing to V9, grouping and drop of PET_G data.

	Regression Coefficient / Area density [kg/m²]	Standard Error [kg/m²]	95% interval
Constant	602,052	27,517	548,118
OTHER_POLYMERS	0,910	0,090	0,733
PVC	1,522	0,059	1,406
PET	1,971	0,066	1,842
CELLULOSICS	1,695	0,034	1,628
PO	0,376	0,013	0,350

Table A2: Summary of bootstrapping results presenting the mean, median, 0.025 and 0.975 quantile for the area densities determined with PO75C and AA106 as well as V9 pre-processing.

	Mean	Median	0.025 quantlile	0.975 quantile
Constant	602.3	602.17	540.7	663.08
OTHER_POLYMERS	0.913	0.912	0.698	1.133
PVC	1.524	1.524	1.344	1.706
PET	1.973	1.973	1.803	2.144
CELLULOSICS	1.694	1.694	1.615	1.775
PO	0.376	0.376	0.345	0.405

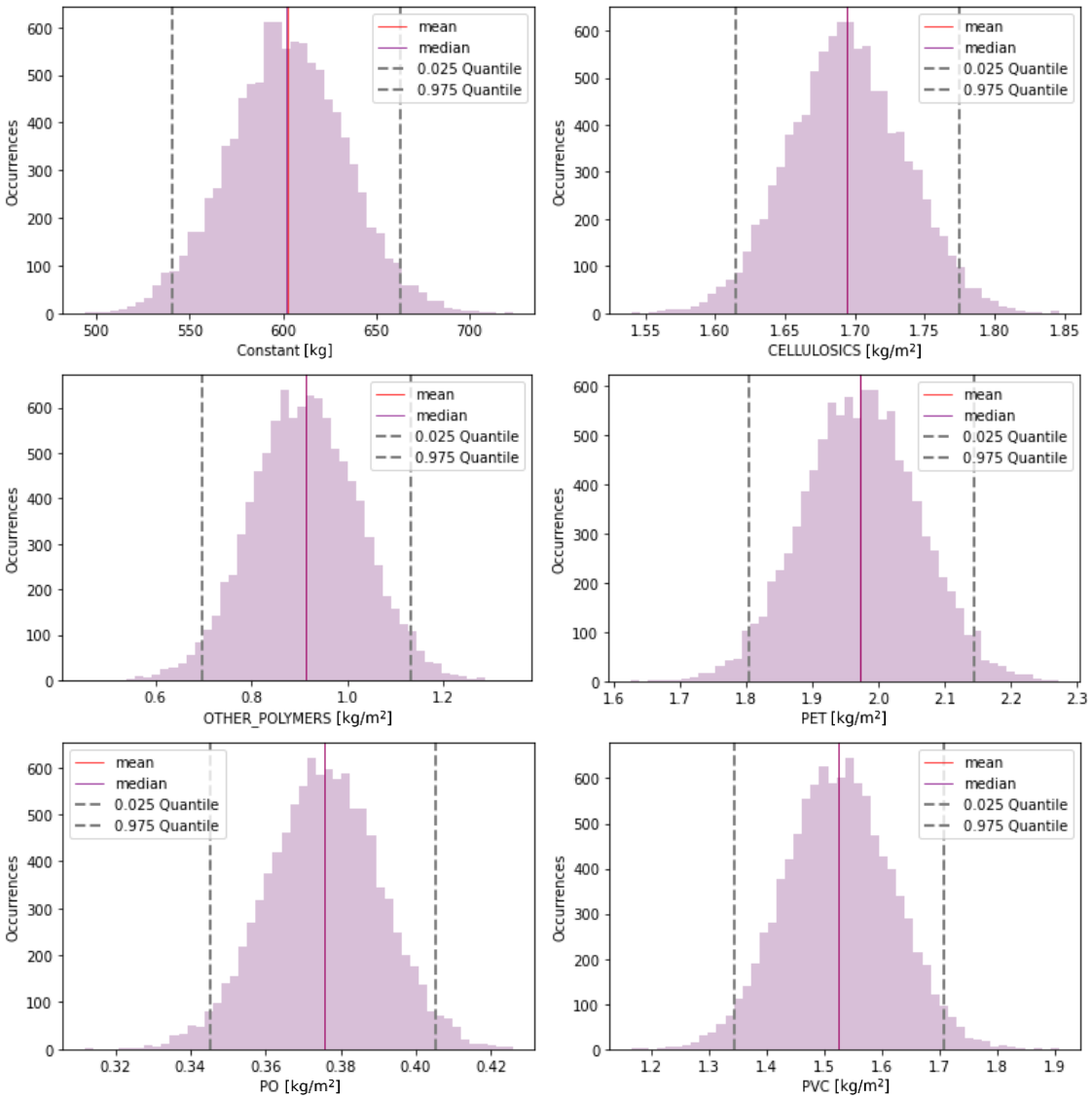

Figure A11: Bootstrapping results for the area densities and the constant of PO75 and AA101 after 10,000 resampling applications and V9 data pre-processing. The mean, the median and the 0.025 and 0.975 quantiles are indicated.

Appendix 9: Area densities with V9 pre-processing and PAPER as distinct category for PO75C and AA101

Below, important figures and information for the area densities determined for PO75C, with belt weigher data from AA101 and V9 data pre-processing as well as PAPER as distinct category, can be found.

Table A3: OLS results for area density prediction with PAPER as distinct category for PO75C using data from PO75C and AA106 after similar data pre-processing to V9, grouping and drop of PET_G data.

	Regression Coefficient / Area density [kg/m²]	Standard Error [kg/m²]	95% interval
Constant	-41,875	15,684	-72,616
BC	0,452	0,066	0,322
OTHER_POLYMERS	0,376	0,025	0,328
PAPER	0,482	0,073	0,338
PVC	0,939	0,061	0,819
PET	3,752	0,114	3,528
PO	0,464	0,008	0,448

Table A4: Summary of bootstrapping results presenting the mean, median, 0.025 and 0.975 quantile for the area densities determined with PO75C and AA106 and V9 pre-processing.

	Mean	Median	0.025 quantlile	0.975 quantile
Constant	-41.9	-42	-76.84	-6.61
BC	0.452	0.451	0.299	0.609
OTHER_POLYMERS	0.376	0.376	0.334	0.418
PAPER	0.483	0.482	0.323	0.645
PVC	0.938	0.938	0.823	1.051
PET	3.754	3.755	3.458	4.043
PO	0.464	0.464	0.444	0.484

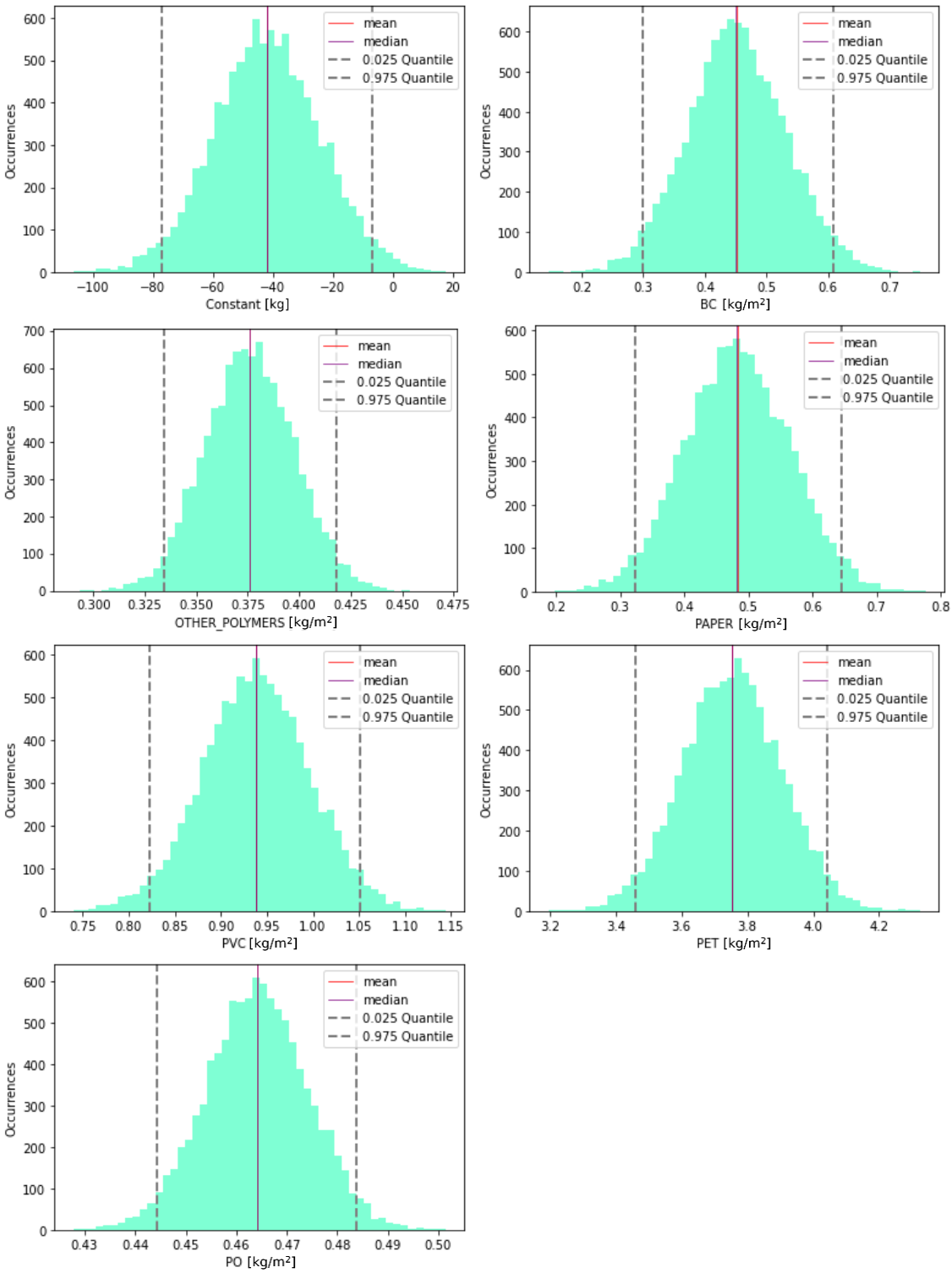

Figure A12: Bootstrapping results for the area densities and the constant of PO75C and AA106 after 10,000 resampling applications and V9 data pre-processing with PAPER as distinct category. The mean, the median and the 0.025 and 0.975 quantiles are indicated.