Internship Report: Business Analyst at Amazon

Anton Zaitsev
University of Luxembourg
0230981826@uni.lu

Contents

Abstract

This report summarizes my internship experience as a Business Analyst at Amazon.com, Inc. (hereafter referred to as Amazon), which took place from March 3rd to August 29th, 2025. The goal of the internship was to learn about Amazon's supply chain operations and contribute to data-driven initiatives aimed at improving network efficiency.

During the internship, I worked on two main projects. The first one involved using non-parametric statistical methods to find backlog levels - a specific metric used in planning - at which sites operate most efficiently. The method used proved effective with sufficient data available but produced unreliable results for more granular data levels. Given the limited time and project not being in the main scope of the internship, I decided to pause this work, noting that it holds potential for future in-depth analysis. The second, core project, called Inbound Maximum Processing Capacity Utilization, consisted of three parts: validation, monitoring, and optimization. For the validation part, I carried out detailed analyses to improve the accuracy of models estimating maximum processing capacities across facilities. These models then served as the foundation for the second part of the project, which focused on identifying over- and under-utilized sites by developing monitoring dashboards, which also allowed to analyze different planning scenarios by varying model input parameters. The final optimization part, aimed at suggesting volume shifts to balance the network, could not be fully completed within the internship timeframe, and was staged for subsequent development.

My contributions had a visible impact across multiple teams within the EU supply chain. Although the scope of the project extended beyond the duration of my internship, I ensured that all work was well-documented, clear, accessible, and laid a solid foundation for continued development and long-term improvement

Keywords: Amazon; supply chain; business analysis; planning; forecasting; modeling; optimization; time series; nonparametric statistics.

1 Introduction

1.1 Background

Amazon operates one of the most complex and successful supply chains in the world, supporting a wide range of business functions from e-commerce to cloud services. According to the latest 2024 letter to stakeholders from Amazon's CEO, Andy Jassy, "Our total revenue grew 11% year-over-year from $575B to $638B" [2]. With more than a million employees, it can be difficult to see how the work of one person, especially an intern, contributes to Amazon's massive business, particularly when the work is not tied to global initiatives. The purpose of this report is to document my internship experience and show how my work provided value to the business and the teams involved.

The role of a business analyst at Amazon covers a wide range of responsibilities. These include developing data extraction and preprocessing pipelines, analyzing data, creating data visualization dashboards for stakeholders, which monitor metrics that are important for real-time decisions, and building optimization algorithms or machine learning models. At Amazon, interns are assigned meaningful projects and are expected to take full ownership of their work. As a result, their contributions are clearly visible and often used by multiple teams across the organization.

During my internship, I worked with databases, created and automated complex data retrieval pipelines, designed clear data visualizations and reports, developed optimization algorithms for business use cases, and conducted analyses using statistical methods. The methods and ideas used to solve business tasks in this report reflect my personal development and highlight the practical value of data-driven approaches in business analysis.

1.2 Problem Statement

The team I worked with needed an analyst skilled in programming languages suited for handling large amounts of data, particularly SQL and Python. Beyond technical skills, it was

important that the analyst could quickly understand new and complex business processes. Although this report focuses specifically on Amazon's supply chain division, where I worked during my internship, the processes involved in supply chain management are often complicated. Fully understanding these processes, knowing how decisions are made and what might cause certain issues, usually requires years of experience.

One key goal of my internship was to use data analysis to better understand these processes. By using data-driven methods, my work aimed to identify the root causes of operational problems and provide insights that could support faster and more accurate planning.

Optimization, approximation, and forecasting also play a significant role in daily operations. As a result, knowledge of relevant techniques and algorithms was essential. It helped me quickly assess which methods and solutions could be applied to a specific use case and which ones were not suitable, depending on the available data and given task.

1.3 Objectives

My goal was to apply my academic background in data science to real-world business problems, while gaining a deeper understanding of large-scale supply chain systems.

During the course of the internship, I worked on several projects with distinct but ultimately interconnected goals. One of the side projects focused on identifying the optimal backlog range for cross-dock sites. The idea was to determine a performance window for a specific metric within which cross-dock operations would run most efficiently. The goal was to avoid underutilizing the site's capabilities while also avoiding overloading the site, which could slow things down and prevent it from processing all the expected volume.

The main project, Inbound Maximum Processing Capacity Utilization, focused on the planned utilization of maximum capacity at cross-dock sites and fulfillment centers. Different planning teams are responsible for estimating site workload volumes and determining the necessary staffing to meet those targets. However, these estimates can sometimes exceed what a site is physically capable of processing, based on various parameters, such as hourly processing rates, product volume shares, etc. My role involved bridging the gap between planning and maximum capacities. First, I was responsible for validating the models used to estimate the maximum processing capacities of the facilities. These capacities were based on historical and planned data, and technical specifications of the machines involved, for example, the processing speed of the conveyors within the product handling system. Second, I developed monitoring dashboards that compared planned processing volumes with modeled capacity limits and enabled the analysis of different planning scenarios. Finally, I prepared the groundwork for an optimization task aimed at distributing volume across the supply chain more efficiently.

While the idea behind these initiatives was similar for both cross-dock sites and fulfillment centers, i.e. modeling maximum capacities and comparing them to planned volumes, the complexity for fulfillment centers was higher. Fulfillment centers operate with a broader set of constraints and a more complicated inbound product flow than cross-dock sites. Developing a similar dashboard for these environments required a more detailed analysis and an account for more parameters. Nevertheless, the goal remained the same: to enable more informed, data-driven planning decisions.

Each of these projects helped improve visibility into how sites were performing, made it easier for different teams to align their plans, and supported more accurate and efficient decision-making. Throughout the internship, I further developed my technical skills in SQL and Python, and gained more experience with data preprocessing and analysis. I learned how to apply these tools in a business environment, which differs from the academic setting.

1.4 Structure of the Report

This report is organized into several sections. After this introductory section, the report will provide an overview of Amazon and its supply chain. It will explain key components of the supply chain, and the roles of cross-dock sites and fulfillment centers. The report will also introduce the main teams involved, including my team and our specific responsibilities.

Following this, the report will focus on the projects carried out during the internship. Each project will be discussed in detail, covering the reasons for undertaking the project, specific tasks performed, results and other relevant details.

The Discussion section will address challenges encountered during the internship and identify potential areas for improvement. Finally, the conclusion section will summarize the work done, discuss possible future directions, and provide reflections on the overall internship experience.

1.5 Terminology

Throughout this report, we repeatedly use several technical and business-related terms. Before moving on to the main part of the report, we want to ensure clarity and readability for the readers. Thus, we have included this dedicated Terminology section. Its purpose is to serve as a quick reference point for those who may need a reminder or clarification of key terms. This section focuses on the most common and essential terms, while the more context-specific terms will be defined at the point of use within the main text.

- Retail: Amazon acts as the primary seller by purchasing products directly from vendors and selling them on its platform.
- FBA (Fulfillment by Amazon): A service model where third-party sellers store their products in Amazon's fulfillment centers. Amazon then handles storage, packaging, and shipping to customers on behalf of the sellers.
- SC (Supply Chain): The end-to-end process involved in purchasing, storing, and delivering products across Amazon's network.
- E2E (End-to-End): Process that covers the entire product flow from the origin point to the final destination.
- FC(s) (Fulfillment Center(s)): Amazon facilities where customer orders are processed, picked, packed, and shipped.
- FC AR Sort: A type of Fulfillment Center equipped with Amazon Robotics (AR), where mobile robots (see Figure 1) carry shelving units (pods) to associates for stowing and picking. Instead of associates walking through aisles to manually retrieve or place items (as is the case in non-robotic, manual sites), the AR system brings the pods directly to designated workstations. This automation reduces walking time, increases efficiency, and allows for more compact storage compared to traditional fulfillment centers.
- Cluster: A group of fulfillment centers located within the same country or geographical region.
- IXD(s) (Inbound Cross-Dock site(s)): Facilities that receive and preprocess vendor or seller products and quickly transfer them to FCs, without long-term storage.
- IB (Inbound): Flow of inventory coming into an Amazon facility. This includes receiving goods from vendors, sellers, or other Amazon sites (such as transhipin flows into FCs).
- OB (Outbound): Flow of inventory leaving a facility. This includes customer order shipments or internal transfers (e.g., transhipout flows from IXDs to FCs or delivery stations).
- TSI (Transship In): The process of receiving inventory at a fulfillment center that has been transferred from another Amazon site (typically an IXD or another FC). It involves unloading, verifying items, and stowing them into inventory.
- TSO (Transship Out): The process of transferring inventory from one facility to another, usually from an IXD to an FC. This involves picking, packing, and loading items for transportation.
- HVE (High Velocity Event): A period of unusually high demand or activity, such as Prime Day or Black Friday, that requires special planning and real-time operational adjustments.

Figure 1: Robots carrying pod shelves for stowing or picking in a robotic FC [1].

- PPT (Production Planning Team): A team responsible for managing capacity planning and ensuring that facilities are able to meet volume demands efficiently.
- S&OP (Sales and Operations Planning): Supply Chain team that deals with Volume and ensures that inventory levels and capacities meet forecasted sales.
- Volume: The number of units processed through a facility. Inbound volume refers to inventory arriving from vendors, sellers, or other facilities; outbound volume refers to units being shipped out to customers or transferred elsewhere in the network.
- Capacity: The maximum amount of volume a facility can handle effectively. This includes: mechanical capacity – the throughput limits of physical equipment; labor capacity – the processing ability based on workforce availability; storage capacity – the amount of inventory space available within the facility.
- TPH (Throughput Per Hour) – Processing speed, representing the number of units handled by a facility within one hour. TPH is often used to evaluate operational efficiency across different processes, such as sortation or stowing TPH.
- UPB (Units Per Bundle): The number of individual items contained within a bundle. Bundles can include containers such as cases, pallets, or totes.
- Pallet: A wooden or plastic platform used to stack and transport multiple cases or totes.
- Case: A box or container that holds multiple individual items of the same product.
- Tote: A reusable plastic container used for organizing, storing, and moving items within Amazon facilities.
- ASIN (Amazon Standard Identification Number): A unique identifier assigned to every product in Amazon's catalog. We often refer to Single ASIN (products of the same kind) or Multi ASIN (a mix of different products) in the context of bundled inventory like pallets or cases.
- NVF (New Vendor Freight): Inbound shipments arriving at IXDs or FCs from vendors or sellers.

- Backlog: The number of customer orders received but not yet fulfilled, which represents pending demand that exceeds the current processing capacity.
- Load Balancing: The process of redistributing TSI between FCs. This is often done to reduce costs or relieve high backlogs at a specific FC.
- Bottleneck: The slowest process in a facility's process flow, which limits the overall throughput and thus the maximum capacity.
- OEE (Overall Equipment Effectiveness): Metric used to measure how efficiently equipment is operating compared to its full potential. It is defined as the product of availability (actual operating time vs. planned production time), performance (actual production speed vs. ideal speed), and quality (good units produced vs. total units started). An OEE score of 100% means perfect production: no downtime, no speed loss, and no defects.
- POC (Point of Contact): In our context, this refers to the individual or team responsible for providing accurate information related to a specific topic or process. For example, a POC might confirm the number of stations available in a warehouse for a given operation, or validate whether a site can realistically process a certain volume (e.g., N pallets per hour).
- ACU (Average Cube per Unit) - Refers to the size of items. Generally used when refering to the fulfillment center inbound and outbound capacities and pic.

2 About Amazon

2.1 Amazon's Supply Chain

Amazon started as an online bookstore. Over time, the company expanded its operations to include a wide range of products and launched an online marketplace that not only sold items directly (retail) but also enabled third-party (3P) sellers to sell their products on the platform (Fulfillment By Amazon). Today, while Amazon operates across several industries, including cloud computing, digital streaming, and logistics, retail and the marketplace remain core components of its business.

These operations function under what is known as the Supply Chain (SC). The supply chain refers to the flow of products from vendors and sellers to customers. A simplified overview of this process is shown in Figure 2. Products can enter the network from vendors (as part of Amazon's retail model) or third-party sellers (under the FBA model). They are received at either cross-dock sites (IXDs) or fulfillment centers (FCs), where inventory is processed and routed accordingly.

IXDs act as hubs where incoming products are sorted and redirected to FCs. At FCs, products are stored until customer orders are placed. This initial movement of goods - from vendors or sellers to FCs - is often referred to as the First Mile. Once an order is placed, items move from FCs to Sortation Centers (Middle Mile), and finally to Delivery Stations (Last Mile), from where they are delivered to customers. This full product journey, from procurement to customer delivery, is referred to as the End-to-End (E2E) pipeline.

Naturally, each stage of this pipeline requires careful planning and forecasting. Capacity models and predictive tools help Amazon anticipate demand, optimize product flows, allocate labor, and manage transportation. While the E2E supply chain includes hundreds of tools and processes, this report will focus on several key components - such as maximum capacity modeling, volume optimization, and transportation cost forecasting - that play a major role in planning volume distribution and workforce allocation.

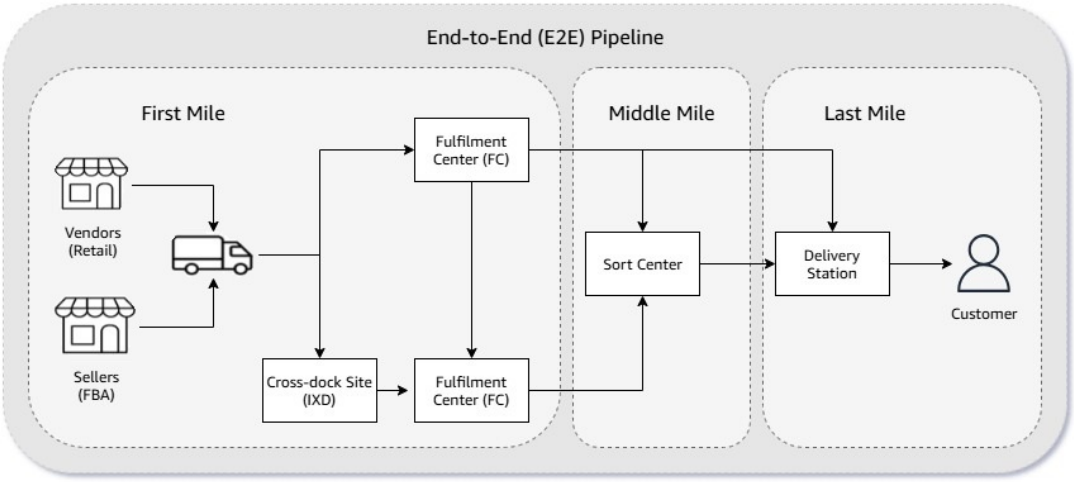


Figure 2: Simplified overview of Amazon's End-to-End (E2E) pipeline. Products are received from either vendors or sellers and then distributed to FCs directly or via IXDs. Then, inventory is shipped from FCs to Sort Centers, from Sort Centers to Delivery Stations and, finally, reaches customers.

2.2 Operational Facilities: Cross-dock Sites and Fulfillment Centers

2.2.1 Cross-dock Sites

Cross-dock sites, or simply IXDs, serve as the initial hubs for distributing incoming inventory across fulfillment centers. Volume that arrives at IXDs from vendors and sellers is called new vendor freight. This volume, i.e. various products, is quickly sorted and prepared for transfer to the appropriate FCs. A detailed overview of this process is shown in Figure 3.

New vendor freight arriving at IXDs come in palletized or fluid-loaded formats. Palletized inventory may contain either a single product (single-ASIN pallet) or multiple products (multi-ASIN pallet). Fluid-loads are always multi-ASIN and therefore require sorting. Single-ASIN pallets do not require processing and are directly assigned to FCs without going through the full IXD workflow, since the whole pallet containts only a single product. This direct path is referred to as the "happy path". The only requirement for happy path inventory is to assign the pallet to an appropriate FC that needs the product.

The sortation process at IXDs is an important component of the supply chain. The more products that can be sorted at this early stage, the less effort is required later in the process. In this context, sortation refers to separating products and assigning them into totes - bins that are later used for stowing products in FCs. Each product is internally mapped to a specific tote, thus the system knows at all times where the products are, making the downstream retrieval process fast and efficient. This early sortation reduces the need for additional handling later in the fulfillment flow, and since IXDs are optimized for high-throughput sortation, they are the preferred location for this step.

However, IXDs have mechanical capacity limits. Some products cannot be processed by sortation machines due to their size, fragility, or weight and must be sorted manually, which introduces additional constraints on throughput rate.

After sortation, the distribution of products from IXDs to FCs is optimized for cost efficiency and inventory demand. Forecasting plays a central role in this process. The system takes into account expected future demand across countries, regions, and cities, and routes products accordingly to ensure fast customer delivery. IXDs are a relatively recent addition to Amazon's network, and innovations such as one-day delivery would not be possible without their role in early-stage sortation and routing.

Products leaving IXDs are either palletized or fluid-loaded into trucks. Palletized shipments

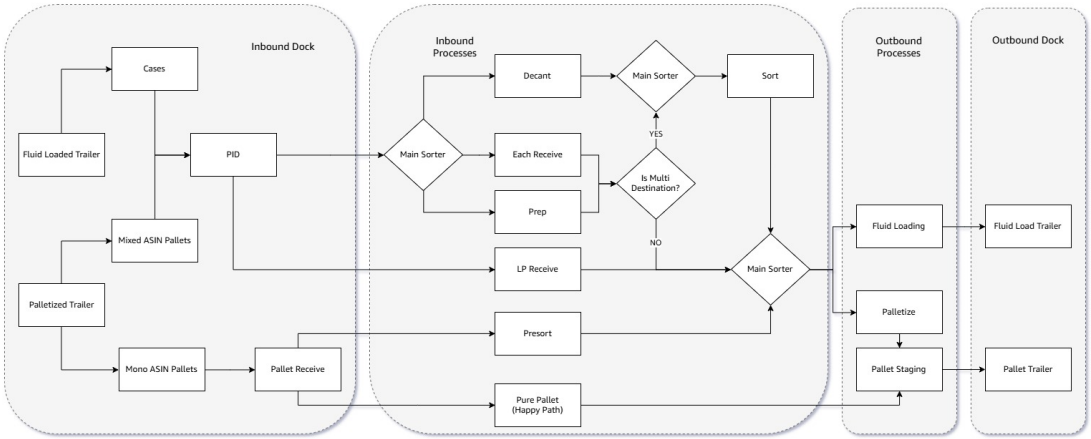


Figure 3: Overview of the IXD product flow. IXDs receive products either as pallets or cases. The pallet receive process varies based on whether the pallet contains multiple products (Mixed ASIN) or a single product (Mono ASIN). Pallets requiring depalletization are routed through the Presort process. For case processing, a key component is the Parcel Identifier (PID) machine, which generates unique IDs for each case and routes them to the appropriate receive stations. Cases are also classified based on whether they are destined for a single or multiple locations. Products with multiple destinations undergo additional sortation before outbound processing. Finally, all inventory is either fluid loaded or palletized for trailer staging and transfer to FCs.

consist of boxes (cases) stacked on a wooden or plastic pallet. Fluid-loading, by contrast, places the product cases directly into trucks, stacking them up to the roof. This method uses space more efficiently and is preferred when possible. After loading, the trucks are dispatched to FCs for the next stage of processing.

2.2.2 Fulfillment Centers

Fulfillment Centers (FCs) are the primary hubs where products are stored and shipped to customers. Inventory that arrives at an FC can come either as new vendor freight or processed from IXDs. It is then unloaded, received into the system, and then stowed, i.e. stored, in storage shelves. Products remain in storage until customer places an order. Once an order is received, the appropriate items are picked from storage, sorted into singles or multis (i.e. orders containing single or multiple products), packed, and prepared for outbound shipment. The final step involves either fluid loading or palletizing shipments for delivery.

The general flow of operations within FCs - how products are received, stowed, and eventually shipped to customers - is illustrated in Figure 4. I had the chance to visit a FC in France, which gave me a good understanding of the processes involved in the FC inventory preprocessing pipeline.

2.3 Capacity

One concept that is important to clarify in the context of operational planning is capacity. In simple terms, capacity refers to the volume of products that a site - such as an IXD or FC - can process within a given time frame. However, the definition and calculation of capacity vary depending on the type of building and the specific operations taking place there.

At IXDs, for example, sortation capacity is determined by mechanical constraints - specifically, how many units the sortation machines can process per hour. These limits are typically based on manufacturer specifications, but are often adjusted using historical data and observed performance. In addition, manual sortation is limited by labor-related factors, such as the number of available sortation stations, their average throughput, the number of working hours per day, and how many shifts are scheduled.

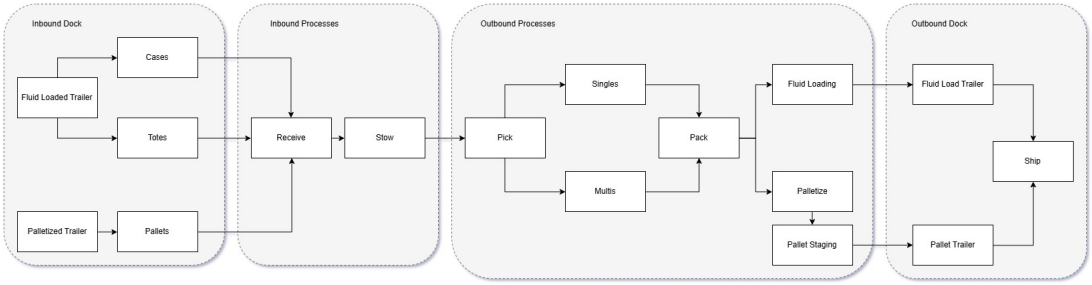


Figure 4: Overview of the FC product flow. FCs receive inventory from IXDs or directly from vendors/sellers in either palletized or fluid-loaded form. Upon arrival, products are received and directed to stow stations, where they are placed into storage bins or pods. When a customer order is placed, products are retrieved from storage (picked), and sent to pack stations. After packing, shipments are prepared for outbound transport, either fluid-loaded directly into trailers or palletized depending on the destination and volume.

Receive capacity is more complex. It depends on a combination of operational and mechanical factors, including the number of truck docks and receive gates, the speed at which associates unload pallets and process incoming cases, the number of associates per shift and the length of those shifts, the composition of inbound inventory (such as how many units arrive per pallet or per case), the mechanical capacity of the receive conveyors, PIDs, and historical throughput performance at the site.

In general, anything that influences how quickly items can be received, sorted, and sent out contributes to the site's overall capacity. Capacity is not a fixed number. It changes based on labor availability, equipment performance, product mix, seasonality, and other operational conditions.

Taking all of these factors into an account and creating a reliable capacity model is a difficult task. Nevertheless, understanding and modeling capacity is critical for planning purposes. If capacity is overestimated, sites may receive more volume than they can process, leading to backlogs and delays. If capacity is underestimated, available resources may be underused. Capacity modeling helps balance volume, labor, and equipment to ensure that sites operate as efficient as possible under always changing business conditions.

3 Teams

Amazon's SC involves many teams that collaborate every day to keep E2E operations running smoothly and efficiently. During my internship, I worked within the High Velocity Events (HVE) team and engaged with several other teams across planning, analytics, and operations.

The HVE team acts as a coordination layer across the entire network during high-demand periods, known as high velocity events. These events are characterized by high increases in customer orders and require planning and monitoring to ensure that the network can handle the extreme volume and that the backlog levels remain within an acceptable range (here, backlog refers to the amount of volume that has arrived at a site but has not yet been processed). Common HVE examples include Prime Week, the holiday season, and Black Friday. Planning for these events begins several months in advance, and the HVE team plays an important role in making sure operational plans are in place. During high velocity events, multiple teams come together to review supply chain performance, compare actual data against planned targets, and coordinate appropriate actions. HVE team also works on validating planning models to ensure that plans and forecasts make sense and align with historical trends.

One of the teams I worked closely with was the ACES Data Analytics team. Their mission is to create and maintain data products to support decision-making across the SC. These products include data dashboards, planning and forecasting models, and datasets. For example, the ACES team developed Maximum Processing Capacity models for FCs and IXDs, which estimates the

maximum volume sites can handle based on machine throughput limits and historical data.

One of the biggest teams in the SC is the Production Planning Team (PPT), which focuses on labor and capacity planning at IXDs and FCs. Since labor is one of the largest variable costs in fulfillment, PPT plays a key role in minimizing costs while ensuring that customer demand is met in a timely manner. The team analyzes trends in human resources metrics such as absenteeism, attrition, and productivity, and uses these inputs to forecast available labor capacity. Based on this forecast, they identify hiring needs or excess capacity for the upcoming quarter. In simpler terms, PPT ensures that each site has the right number of workers at the right time.

Another important group is the Sales and Operations Planning (S&OP) team. Their role is to collect inventory flow forecasts from retail and FBA teams, as well as partner forecasts for inbound and outbound volumes - that is, the volume entering and exiting a site - and translate these into an operational plan. This plan outlines the expected flow of inventory into and out of each node (i.e., site) in the network, including vendor shipments, customer returns, IXD to FC transfers, and more. The plan is updated weekly and typically covers a 16-week horizon. In essence, S&OP is responsible for short- and long-term demand and volume forecasting across the network.

I also interacted with the IBET and OBET teams, which are part of the daily execution group. IBET handles inbound volume forecasts, while OBET focuses on outbound flows. They ensure that short-term changes in demand are reflected in daily plans. While some teams manage operations at the level of individual FCs or IXDs, others focus on broader scopes such as country-level or cluster-level planning (a cluster refers to a group of FCs located within the same region or country).

These were the main teams I worked with during my internship. Naturally, there are many other teams involved in the SC, such as Transportation teams that handle truck routing and scheduling, or Science teams that develop artificial intelligence solutions for various problems.

4 Project 1: Optimal Backlog Range Identification

4.1 Introduction

Finding the right backlog level for Amazon facilities has been an ongoing challenge, and several teams have explored different approaches. As a side project during available downtime, our team also decided to work on this problem, aiming to identify the optimal backlog range in which a site can operate efficiently, without being under- or overutilized. For example, the IBET team approached this by estimating a safety threshold based on two years of historical and planned backlog data. They modeled the relative error between the two using a kernel density estimation (KDE) technique, then ran Monte Carlo simulations across various backlog levels. For each level, they calculated the probability of breaching the maximum holding capacity of the site, that is, how much backlog a site can physically hold in its inbound area. The threshold was defined as the highest level of backlog that keeps this risk within an acceptable range. In other words, their goal was to find the maximum backlog that each site could hold before breaching its capacity limits. In contrast, our approach focused on finding an optimal backlog that a site should operate at by directly analyzing the historical relationship between vendor receipts and backlog levels.

4.2 Hypothesis

To approach this task, we initially considered using correlation analysis to find the optimal backlog range. Our assumption was that at low backlog levels, an increase in receipts would lead to an increase in backlog, resulting in a positive correlation. At higher backlog levels, the site's processing capacity, limited by staffing and equipment, becomes the bottleneck. Beyond this point, receipts no longer increase proportionally with backlog, and the correlation weakens. Based on this, we formulated two hypotheses on how the relationship between backlog and receipts looks like: one in which the relationship resembles a downward-facing parabola, and

another in which receipts rise rapidly but then plateau once the maximum capacity is reached (see Figure 5).

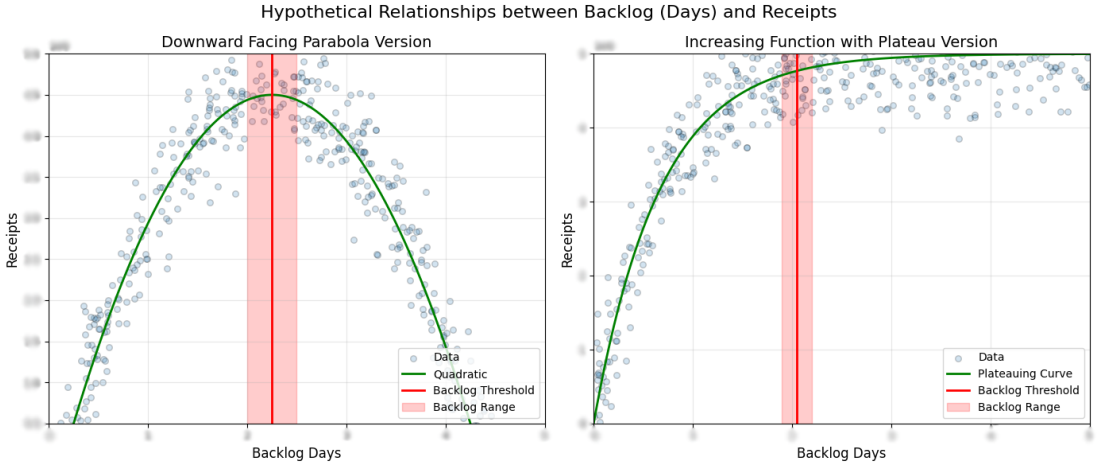


Figure 5: Two hypothetical relationships between backlog days and receipts. The first curve (left) shows a downward-facing parabola: receipts increase with backlog until reaching a peak, then decline. The second curve (right) shows an increasing function that rises steeply at low backlog levels, but plateaus once site capacity is reached. In both cases, the red-shaded "Backlog Range" around the start of the plateau or peak represents the hypothetical optimal backlog window, where site capacity is used efficiently without being underutilized or overutilized.

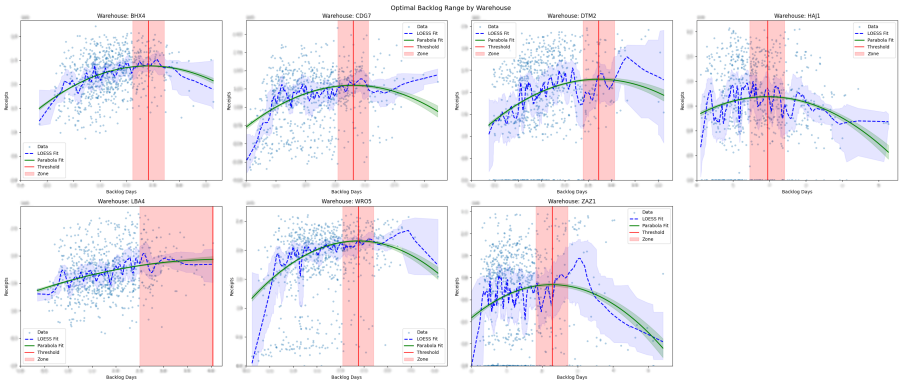
4.3 Methodology

To test these hypotheses, we first prepared the dataset by removing outliers and minimizing noise. Specifically, we applied a quantile-based filter: for each warehouse and weekday group, we kept only the data between the 5th and 95th percentiles of receipts. This step ensured that extreme values did not distort the analysis and helped us focus on representative operational behavior.

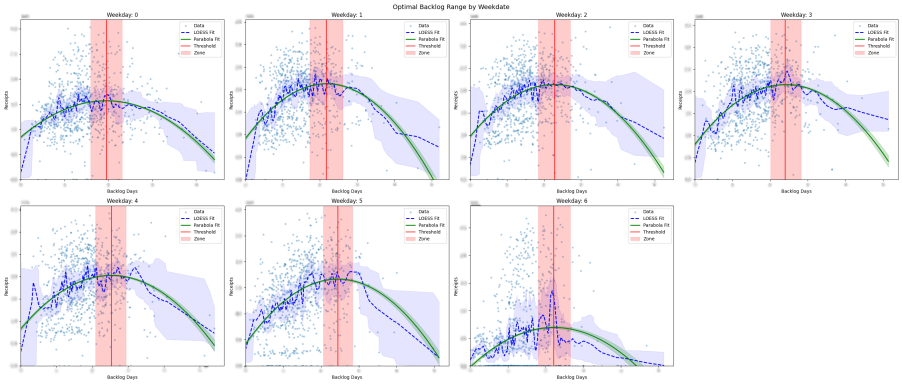
Then, we initially considered using a correlation-based method to estimate optimal backlog levels. One idea was to divide the backlog range into fixed intervals (e.g., $[0, 0.5], [0.5, 1]$, etc.) and identify where the correlation between backlog and receipts shifted from positive to negative. However, this approach seemed problematic. It was highly sensitive to how the intervals were defined and susceptible to noise, making it difficult to interpret and prone to overfitting. We decided not to pursue this approach, as it raised too many open questions and would likely lead to time-consuming fine-tuning without clear benefit.

Next, we explored fitting a quadratic function to the receipts versus backlog relationship, expecting to see a downward-facing parabola, as per our hypothesis. While this worked in some cases, the fit did not always produce a reliable peak, from which we could deduce an optimal backlog range. We then tested a non-parametric regression approach using LOESS (Locally Estimated Scatterplot Smoothing). However, this method presented two challenges: when using a high LOESS fraction parameter, the fit became too smooth and essentially mimicked a quadratic curve we achieved earlier. When using a low LOESS fraction parameter, the fit better captured local patterns in the data, but the resulting curve was too irregular ("bumpy"). This made it difficult to easily identify a consistent and reliable method to determine the optimal range of backlog values.

Facing these challenges, we adopted a combined approach. We first applied LOESS smoothing with relatively low fraction parameter value to reduce local noise and capture the overall trend between backlog and receipts. We then fitted a quadratic function to the smoothed LOESS curve, which allowed us to extract a single peak value, representing the backlog level beyond which additional workload led to reduced processing efficiency.



(a) Optimal backlog ranges per warehouse. Each subplot shows the relationship between backlog days and receipts for a given warehouse.



(b) Optimal backlog ranges per weekday. Each subplot shows the backlog–receipts relationship for a specific weekday (0 = Monday, 6 = Sunday).

Figure 6: Comparison of optimal backlog ranges across warehouse (top) and weekday (bottom) dimensions. The LOESS smoothing (blue dashed) captures the local trend, while the quadratic fit (green) helps estimate the backlog threshold (peak of the parabola, red vertical line) and define the optimal operational zone (red transparent zone around the peak). Semi-transparent blue and green zones correspond to 95% confidence intervals for LOESS smoothing and quadratic fit, respectively, calculated using bootstrap resampling. These analyses are used to identify IXD operating zones that balance throughput and system efficiency.

To estimate the optimal backlog range, rather than just a single point, we first calculated the slope of the quadratic fit, then normalized it to the $[0, 1]$ range, and finally identified intervals where the slope remained below a defined threshold. This range was interpreted as the "optimal zone" where the site operated efficiently, i.e., avoiding both underutilization and overloading. We used a slope threshold of 0.15, which typically corresponded to a zone of approximately 0.5 backlog days.

Knowing that operational behavior varies by both warehouse and weekday, we developed four analyses. First, we examined the global relationship between backlog and receipts at the warehouse level (Figure 6a). Second, we repeated the analysis at the weekday level to explore day-specific patterns (Figure 6b). Third, we applied the method jointly across both dimensions by analyzing each warehouse–weekday combination (Figure 22). However, due to limited data in some of these subgroups, not all combinations yielded stable results. To mitigate the issue of data sparsity, we conducted a final intersection analysis. We took the previously estimated ranges from the warehouse- and weekday-level models and computed their intersection for each warehouse–weekday pair. If the intersection was non-empty, we defined it as the recommended operational zone for that combination (see Figure 21).

We also tested another approach, where we would combine LOESS and binned-averages techniques, however we decided to ultimately switch to this method, as it results in the most easily interpretable results.

4.4 Presenting Analysis

After completing the main analysis, we presented our approach to the IBET team and their IXD expert. They reviewed the estimated backlog ranges and confirmed that the results were reasonable and aligned with their operational expectations. They appreciated the simplicity and interpretability of the approach, but also highlighted two key considerations. First, our method did not incorporate a safety mechanism, for example accounting for the risk of breaching the site's maximum holding capacity. Second, they pointed out that the ranges we calculated could be useful to planners when deciding how much volume to route to a site, helping avoid both over- and under-utilization.

We agreed with both observations and proposed a hybrid strategy: we would use their modeled safety thresholds as upper cutoffs for the ranges we identified. For instance, if our analysis produced an optimal backlog range of $[1.5, 2.3]$ days for a given site and weekday, but their maximum holding capacity threshold was 2.1 days, we would adjust our recommendation to $[1.5, 2.1]$.

To further validate the flexibility and generalizability of our method, they asked us to extend the analysis to FC AR sites, using backlog in units instead of days, and to segment the analysis by quarter. This final request reflected a mutual understanding that just as optimal backlog thresholds may vary by weekday due to staffing and planning dynamics, they can also vary seasonally, particularly during high-volume periods like Q4.

4.5 Results

The effectiveness of our method was closely tied to the volume of data available. We observed that increasing the level of feature segmentation - such as grouping by quarter in addition to warehouse and weekday - often led to unstable or unreliable results. This was the primary reason we chose to use the intersection of warehouse-level and weekday-level ranges.

While splitting the data by quarter did yield promising results for some site–quarter combinations, the outputs were overall inconsistent. In several cases, the fitted curves contradicted our hypothesis: some showed a continuous decline, some rose steadily, and others formed concave parabolas that lacked clear interpretation (see Figure 7). These patterns suggest that the method, in its current form, may not generalize well to all segments without further refinement.

Due to limited time, we did not pursue deeper investigation or adjust the strategy for these cases. However, we believe the approach has potential if paired with more robust data preprocessing and additional filtering rules to handle edge cases more reliably.

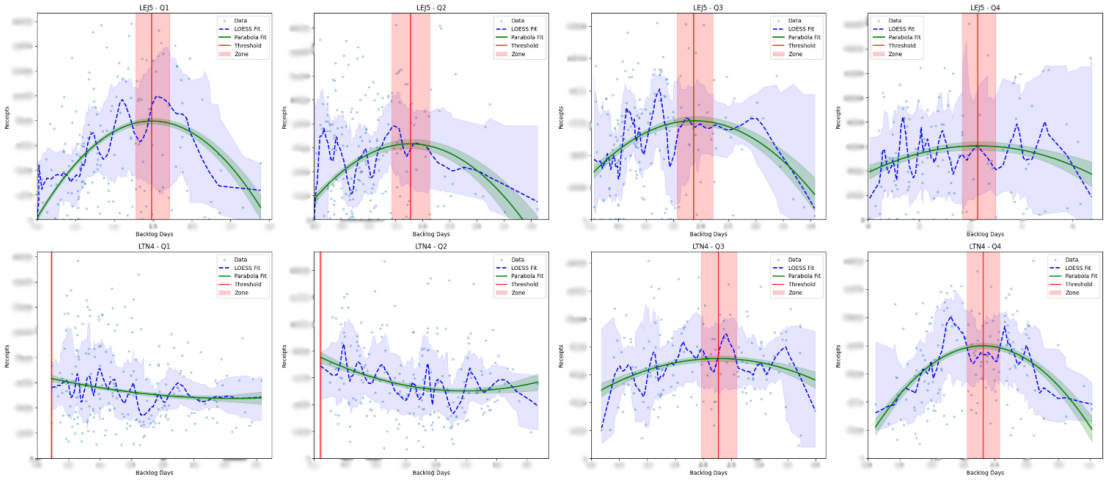


Figure 7: Backlog analysis results for LEJ5 and LTN4 FC robotic sites, using Q1–Q4 data split. While the results for LEJ5 align well with expectations, LTN4 produced unreliable results in Q1 and Q2, illustrating the limitations of the approach under low or noisy data conditions.

5 Project 2: Inbound Maximum Processing Capacity Utilization

The ultimate goal of the Inbound Max Capacity Utilization project was to optimize the flow of the planned volume across the whole EU supply chain network and better align planned volumes with maximum processing capacity limits. The idea was to access each site's maximum processing capacity, compare it against planned volume, and then identify imbalances and bottlenecks. If the planned volume of a facility is significantly below its maximum capacity, the model would flag it as underutilized, suggesting that more volume can be routed there. In contrast, if the planned volume exceeded the capacity, it would be flagged as over-utilized, recommending to redirect excess volume elsewhere. In this way, the project aimed to suggest a more balanced and efficient picture of the inbound supply chain network.

We had three objectives. First, since the project relied on maximum processing capacity models developed by another team, the goal was to validate these models and ensure their reliability. The second part focused on creating monitoring dashboards to compare forecasted volumes with modeled capacity limits and to enable the analysis of alternative planning scenarios. The final objective was to lay the foundation for optimizing the volume across the EU supply chain. The project was structured in two parts: the first focusing on cross-dock sites and the second on fulfillment centers.

5.1 IXD Maximum Processing Capacity Utilization

5.1.1 Introduction

The Inbound Maximum Processing Capacity Utilization project began with IXDs. It was initiated to address a key limitation in IXD network planning: the lack of reliable data on what sites can realistically process. Planning teams faced risks of under- or over-utilizing available resources, since they mostly only relied on historical site throughput. These risks were especially pronounced during peak periods, where inbound volumes and demand drastically increase.

The goal of this work was to let planners see by how much the site is planned to be utilized versus the modeled maximum capacity, giving visibility into total capacity utilization, sortation utilization, case and pallet utilization, and including the utilization of all the subprocesses involved in the IXD process flow. By comparing what is planned with what is realistically possible, teams can quickly identify whether a specific volume can be efficiently processed at an IXD. In cases where over-utilization was identified, planned volumes could be reallocated to other IXD sites with available capacity, ensuring a more balanced and efficient use of the network.

The project was structured in several phases. First, our team proposed the initiative to various stakeholders, including PPT and S&OP teams, from which, after approval, we started gathering customer requirements, that is, what they wanted to see as the output of our work. Next, my work focused on validating the maximum processing capacity models developed by the ACES team. This validation involved assessing model inputs and outputs, comparing the modeled values against historical data, and ensuring full inputs traceability, so that our reported numbers are justifiable. Then, the task was to develop a monitoring dashboard that shows how much of sites' maximum processing capacities are planned to be utilized, based on the planned values and modeled maximum processing capacity values.

5.1.2 ACES Maximum Processing Capacity Model

The Maximum Processing Capacity (MPC) model was developed by the ACES team for two scenarios: forecast and long-term. The forecast scenario uses planned and forecasted input data, such as expected hourly rates for specific subprocesses like receiving or stowing, to estimate capacity in the upcoming weeks. The long-term scenario, which we validated, is based on observed operational data from peak weeks in 2024 together with forecasted peak 2025 data, making it suitable for testing the model logic and ensuring robustness under high-demand conditions.

For the long-term scenario, site limits are defined using the 85th percentile of historical through-

put rather than theoretical machine maximums, reflecting a sustainable and realistic performance level. For example, while a machine might be capable of achieving a certain maximum rate under ideal conditions, the 85th percentile represents a level that is both ambitious and achievable during peak periods.

In both scenarios, the model combines multiple data sources, including historical throughput, machine specifications, units-per-bundle values, and detailed process configurations. It also requires a full mapping of the end-to-end process flow, including all capacity funnels and subprocesses, to accurately estimate overall capacity and identify bottlenecks (see Table 1).

Table 1: Historical units per bundle values used in the IXD MPC model (values anonymized for confidentiality). These figures are based on peak weeks in 2024 and serve as key inputs for calculating maximum processing capacity.

Inventory Type	**Units/Case (Fluid)**	**Units/Case (Pallet)**	**Units/Pallet**
FBA	X.X	X.X	X.X
Retail	X.X	X.X	X.X

Note: 'X.X' indicates average units per bundle from 2024 peak weeks.

In addition to this array of inputs, the model incorporates the number of assets (i.e., machines) per process, processing rates per asset, labor availability (via shift length). The model differentiates between multiple process types, such as manual unloading, conveyable lines, and sortation systems, to simulate the end-to-end capacity pipeline and identify bottleneck processes. Refer to Table 2 to see one part of the model outputs, which shows maximum machine capacities of different processes involved in the IXD pipeline.

Table 2: Isolated process step capacity outputs from the ACES IXD MPC model (most of the data is hidden due to confidentiality). The table presents modeled maximum daily processing capacities and utilization levels for key IXD processes. These modeled values are used to identify potential bottlenecks (highlighting which processes are operating near full capacity and which have available capacity) as well as to calculate overall maximum throughput.

Process Name	**Container Type**	**UPB**	**Daily Working Hours**	**Daily Max Capacity**	**Asset Utilization %**
Sort	Each	X.X	X.X	X.X	X.X
PID	Case	X.X	X.X	X.X	X.X
Manual Unloading	Pallet	X.X	X.X	X.X	X.X
Inbound Dock Space	Pallet	X.X	X.X	X.X	X.X
Fluid Loading	Mixed	X.X	X.X	X.X	X.X
Fluid Unloading	Case	X.X	X.X	X.X	X.X
Truck Loading	Pallet	X.X	X.X	X.X	X.X
Corral Conveyor	Case	X.X	X.X	X.X	X.X
Outbound Dock Space	Pallet	X.X	X.X	X.X	X.X

Note: Additional processes and modeled features are omitted for confidentiality.

Finally, after retrieving historical UPB values, product mix shares, machine capacities, and modeling maximum process capacities, keeping in mind the end-to-end process flow, ACES MPC model produces a set of final output metrics for each site. These include the estimated daily and weekly maximum processing capacities, the primary bottleneck process, and share breakdowns between sort and non-sort volume.

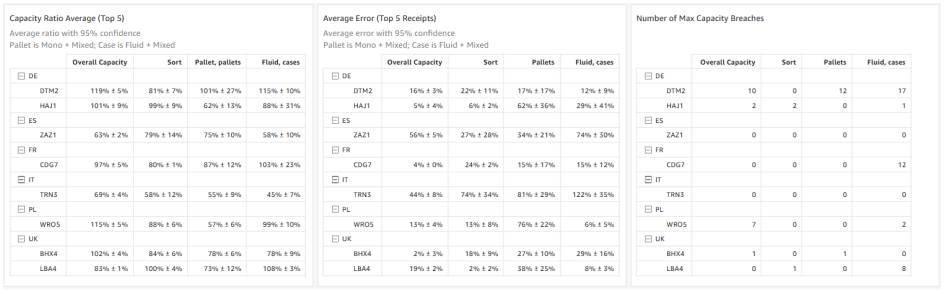
The math behind these calculations is relatively simple. However, the logic is complex. The model needs to account for many interdependent constraints: how volumes flow through different process steps, how capacity is affected by product configuration (case vs. pallet), and where bottlenecks may appear in the IXD processing pipeline. The final output of the MPC model depends on two adjustable parameters: pallet share, which asks out of all inbound volume how much is in pallet, and perfect pallet share. Perfect pallet share represents pallets that flow directly to IXD outbound, thus not affecting the total inbound capacity (only inbound dock capacity, which is usually not a bottleneck). For example, if the planned volume includes 5 million Fluid Case units, 2 million Mixed ASIN Pallet units, and 1 million Mono ASIN Pallet

units—with 10% of the Mono ASIN Pallets being perfect pallets—the adjusted volume becomes: $5M + 2M + 1M - (1M \times 0.1) = 7.9M$. This adjustment lowers the calculated utilization and provides a more realistic estimate of required capacity.

5.1.3 Validation Analysis

Since our work relied heavily on the outputs of the MPC model that we did not develop ourselves, we needed to be confident that those outputs were trustworthy. To do that, we performed a simple validation exercise: we compared the estimated capacities of the model with what actually was received and processed for each IXD.

We focused our analysis on the peak period of 2024, as it represents the time when warehouses operate at or near their maximum processing capacity. In some cases we limited the analysis to the top five inbound days within the selected period (for example when computing the average relative error), which is logical given that the main purpose of the model is to estimate the true maximum capacity of a site. We based our validation on three main points. First, we checked the average utilization for each receive process in the selected period (where utilization is defined as the ratio of the historical volume and the modeled maximum). This covered total receipts, sortation, pallet receipts, and fluid receipts (see Figure 8, left view). Note that metrics involving 95% confidence intervals, such as relative error or mean breach percentage, were computed using $z-$scores when the number of data points was 30 or more ($z = 1.96$), and $t-$distribution values when fewer data points were available (e.g., $t = 2.064$ for 25 points, $t = 2.093$ for 20 points).



Capacity Ratio Average (Top 5)

Average ratio with 95% confidence
Pallet is Mono + Mixed; Case is Fluid + Mixed

	Overall Capacity	Sort	Pallet, pallets	Fluid, cases
DE				
DTM2	119% ± 5%	81% ± 7%	101% ± 27%	115% ± 10%
HAJ1	101% ± 9%	99% ± 9%	62% ± 13%	88% ± 31%
ES				
ZAZ1	63% ± 2%	79% ± 14%	75% ± 10%	58% ± 10%
FR				
CDG7	97% ± 5%	80% ± 1%	87% ± 12%	103% ± 23%
IT				
TRN3	69% ± 4%	58% ± 12%	55% ± 9%	45% ± 7%
PL				
WRO5	115% ± 5%	88% ± 6%	57% ± 6%	99% ± 10%
UK				
BHX4	102% ± 4%	84% ± 6%	78% ± 6%	78% ± 9%
LBA4	83% ± 1%	100% ± 4%	73% ± 12%	108% ± 3%

Average Error (Top 5 Receipts)

Average error with 95% confidence
Pallet is Mono + Mixed; Case is Fluid + Mixed

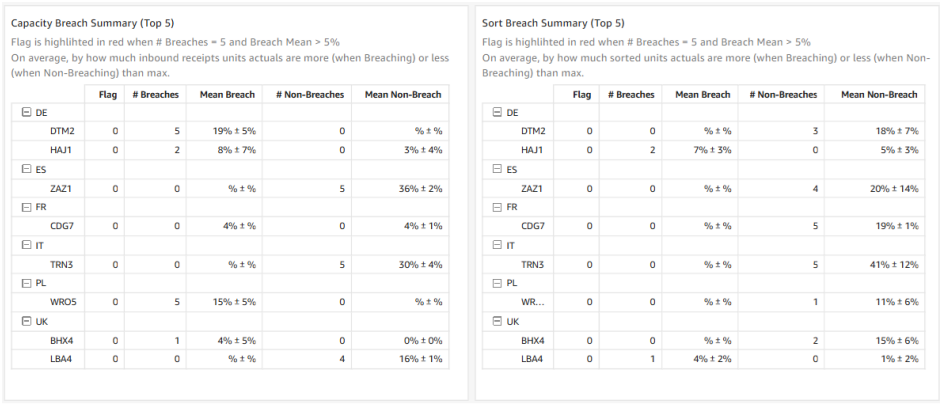
	Overall Capacity	Sort	Pallets	Fluid, cases
DE				
DTM2	16% ± 3%	22% ± 11%	17% ± 17%	12% ± 9%
HAJ1	5% ± 4%	6% ± 2%	62% ± 36%	29% ± 41%
ES				
ZAZ1	56% ± 5%	27% ± 28%	34% ± 21%	74% ± 30%
FR				
CDG7	4% ± 0%	24% ± 2%	15% ± 17%	15% ± 12%
IT				
TRN3	44% ± 8%	74% ± 34%	81% ± 29%	122% ± 35%
PL				
WRO5	13% ± 4%	13% ± 8%	76% ± 22%	6% ± 5%
UK				
BHX4	2% ± 3%	18% ± 9%	27% ± 10%	29% ± 16%
LBA4	19% ± 2%	2% ± 2%	38% ± 25%	8% ± 3%

Number of Max Capacity Breaches

	Overall Capacity	Sort	Pallets	Fluid, cases
DE				
DTM2	10	0	12	17
HAJ1	2	2	0	1
ES				
ZAZ1	0	0	0	0
FR				
CDG7	0	0	0	12
IT				
TRN3	0	0	0	0
PL				
WRO5	7	0	0	2
UK				
BHX4	1	0	1	0
LBA4	0	1	0	8

Figure 8: Summary view which includes the number of breaches for each process type across the selected period (right), average top 5 inbound volume days historical-to-modeled utilization ratios (left), and average relative error during top 5 inbound volume days (middle).

Second, we examined how many breaches occurred (days where historical volumes exceeded the modeled maximum) and reported the mean breach percentage with a 95% confidence interval, along with the same statistics for non-breaches (see Figure 9).



Capacity Breach Summary (Top 5)

Flag is highlihted in red when # Breaches = 5 and Breach Mean > 5%
On average, by how much inbound receipts units actuals are more (when Breaching) or less (when Non-Breaching) than max.

	Flag	# Breaches	Mean Breach	# Non-Breaches	Mean Non-Breach
DE					
DTM2	0	5	19% ± 5%	0	% ± %
HAJ1	0	2	8% ± 7%	0	3% ± 4%
ES					
ZAZ1	0	0	% ± %	5	36% ± 2%
FR					
CDG7	0	0	4% ± %	0	4% ± 1%
IT					
TRN3	0	0	% ± %	5	30% ± 4%
PL					
WRO5	0	5	15% ± 5%	0	% ± %
UK					
BHX4	0	1	4% ± 5%	0	0% ± 0%
LBA4	0	0	% ± %	4	16% ± 1%

Sort Breach Summary (Top 5)

Flag is highlihted in red when # Breaches = 5 and Breach Mean > 5%
On average, by how much sorted units actuals are more (when Breaching) or less (when Non-Breaching) than max.

	Flag	# Breaches	Mean Breach	# Non-Breaches	Mean Non-Breach
DE					
DTM2	0	0	% ± %	3	18% ± 7%
HAJ1	0	2	7% ± 3%	0	5% ± 3%
ES					
ZAZ1	0	0	% ± %	4	20% ± 14%
FR					
CDG7	0	0	% ± %	5	19% ± 1%
IT					
TRN3	0	0	% ± %	5	41% ± 12%
PL					
WR...	0	0	% ± %	1	11% ± 6%
UK					
BHX4	0	0	% ± %	2	15% ± 6%
LBA4	0	1	4% ± 2%	0	1% ± 2%

Figure 9: Number of breaches and non-breaches, with mean breach percentage and 95% confidence intervals (only shown for receipts and sortation).

Finally, we compared the daily 95th percentile values within the selected period against the modeled maximums (see Figure 10). We believe that this approach covered the full and relevant range of validation checks needed to assess the reliability of the model.

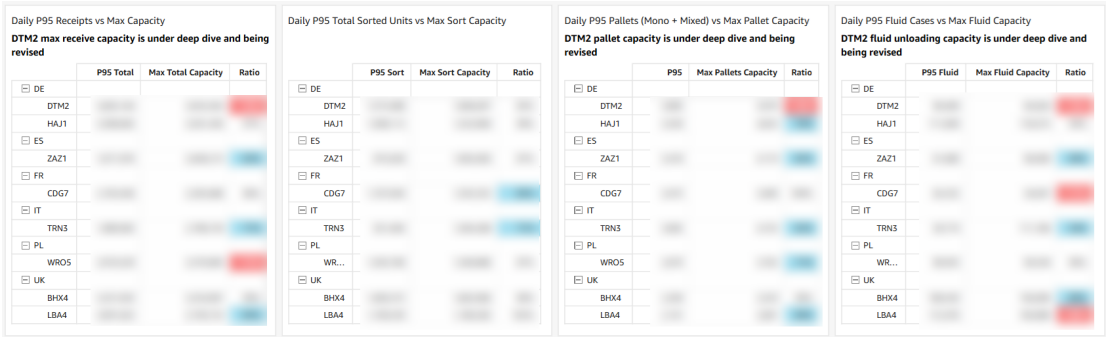


Figure 10: Comparison of daily 95th percentile values, here for total receipts, sortation, pallets, and fluid units against the modeled maximum capacities for each site. Light blue cells indicate utilization below 85% and light red cells indicate utilization above 105%, highlighting potential underestimation or overestimation in the modeled maximum capacity.

Thanks to the validation analysis, we were able to identify several issues in the MPC model. Our focus was primarily on total capacity, pallet capacity, and sortation capacity, particularly in cases where utilization fell outside the 85%–105% range. We selected this range as optimal because the modeled capacities depend on many factors, including the product mix and units per pallet or case. The long-term MPC model uses product mix forecasts from the S&OP team for peak 2025, while the units per bundle values are based on peak 2024. Naturally, these assumptions do not match the historical data used for comparison. Additionally, we consider it more acceptable for the model to slightly overestimate capacity than to underestimate it. This is why the lower bound (85%) is set further from 100% than the upper bound (105%). While overestimating capacity is not ideal, underestimating it poses higher risks, such as inaccurate planning, missed or delayed deliveries, customer dissatisfaction, and reduced operational efficiency.

When we found that certain modeled values deviated too far from historical data, we investigated further to understand the root cause of the discrepancies. For sortation, we broke the process down into its key subprocesses: manual sortation, UIS 20LB, and UIS 5LB. We reviewed the assumed hourly rates and the number of stations used in the model. In cases where the hourly rate assumptions were incorrect, we flagged them to the ACES team. In other cases, the rates were reasonable, but the assumed number of stations was not aligned with actual site configurations. We asked the ACES team to validate those figures directly with the sites. As a result of this work, we significantly reduced the relative model error. The error for total sortation dropped by 18 percent, for UIS 20LB by 38 percent, for UIS 5LB by 26 percent, and for manual sortation by 23 percent. For total site capacity, the final average relative error now stands at 7 percent. This excludes the TRN3 site, which is a newly opened location with limited and unreliable data. If we include TRN3, the error increases to 10 percent.

Pallet capacity was more difficult to validate. Surprisingly, data on the number of processed pallets and processed pallet units was not available and had not been tracked in any system. We had a long series of conversations with different teams and IXD experts to determine the best available data to use. After investigating multiple dashboards and sources, we ultimately decided to use vendor data. When vendors ship products to Amazon warehouses, they must declare the number of units and pallets. Although this data is not perfect, since it does not represent the number of received and processed pallets, we decided to use it, but with caution.

5.1.4 Retrieving and Preprocessing Data

After the validation analysis we moved on to the development of the utilization dashboard. All the dashboards were developed using Amazon QuickSight [3], a cloud business intelligence service

that allows users to build interactive dashboards and visualizations directly on AWS. The goal was to combine planned capacities with outputs from the ACES MPC model to provide planners with a view of planned site utilization from weeks +1 to +16. First, we needed to define a proper and optimized way to retrieve all necessary data. Since the data was not available from a single source, and the QuickSight tool does not support combining visuals from different datasets or databases on the same view, we had to build an intermediate data layer.

The solution was to extract and preprocess data separately (directly in the query) from each database, ensuring that all required components (e.g., planned & forecasted receipts, modeled maximums, historical UPBs and receipts) were pulled into unified datasets. These datasets were then joined on common keys such as site name and week number to create a single dataset instance that could be consumed in QuickSight. See Figure 11 for an overview of the complete data pipeline.

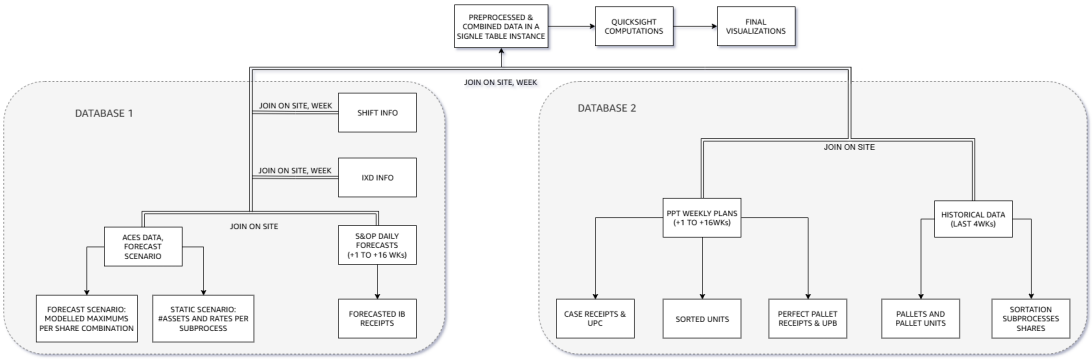


Figure 11: Data architecture behind the IXD Maximum Processing Capacity Utilization dashboard. The pipeline integrates four key sources: (1) ACES-modeled MPC data (2) S&OP daily forecasts (3) PPT weekly plans; (4) historical data, used to estimate case/pallet shares and subprocess breakdowns not directly available in plans. These sources are joined and transformed into a unified dataset used for utilization calculations in QuickSight.

The final joined dataset includes forecasted and planned data (from S&OP and PPT sources), modeled MPC outputs (from ACES-provided tables), historical receipts (from the past four weeks), and supporting metadata such as IXD site-level configurations. Since the plans and forecasts did not contain information on the total number of pallets or pallet units, we retrieved this detail from historical data and applied the historical shares to the forecasted total units to estimate how many units would likely arrive on pallets. Although developing a dedicated time-series forecasting model for this task could have produced more precise results than using a simple four-week average, we did not have sufficient bandwidth within the project timeline to implement it. All required transformations and joins were completed in advance using SQL-based preprocessing during data extraction. This approach enabled us to deliver a single, unified table that could be seamlessly used in the QuickSight dashboard.

We relied on both historical data and S&OP forecasts because the PPT plans alone do not provide a complete view of planned receipts. Specifically, PPT does not include the total planned inbound units or their breakdown into cases and pallets. To fill this gap, we used the total received units from S&OP forecasts and derived the case and pallet shares from historical data covering the past four weeks. It is important to note that the ACES forecast model also depends on assumed pallet and perfect pallet shares. For this reason, we used the historical pallet share along with a user-input control for the perfect pallet share to map the planned receipts to the correct configuration used in the MPC model.

5.1.5 Utilization Dashboard

Once we successfully preprocessed and merged all the necessary data into a single dataset, we began developing the dashboard. Based on stakeholder feedback, we structured the dashboard

into two views. Each view included a main section displaying utilization percentages, and deep dive section that showed planned units, modeled maximum capacities, and other relevant data. The first view focused on daily capacity utilization, specifically for total capacity and sortation capacity (see Figures 12 and 13).

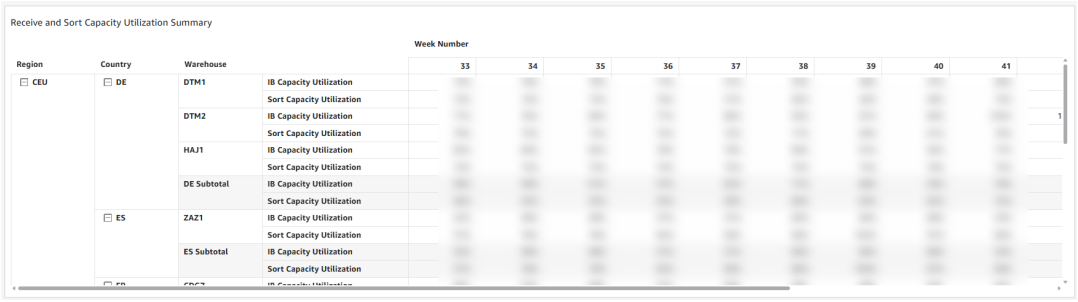


Figure 12: Planned maximum processing capacity utilization for total and sortation capacity. The view also includes planned sortation subprocesses, such as UIS 20LB, UIS 5LB, and manual sortation.

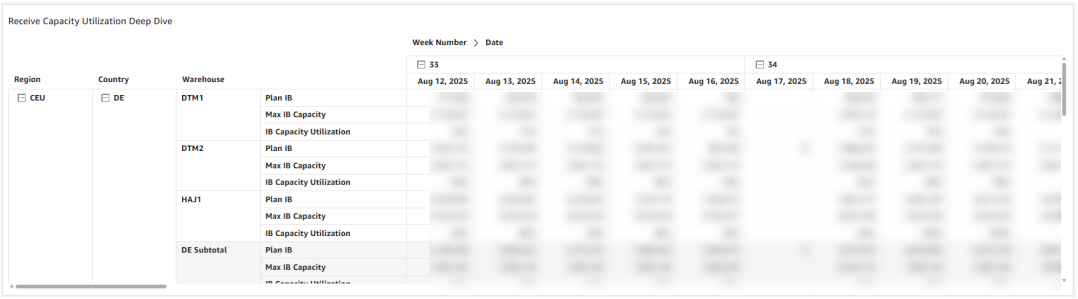


Figure 13: Deep dive view showing weekly planned total receive and sortation units, and derived utilization values.

The second view focuses on pallet, depalletization, case and fluid case receive capacity utilization. Similar to the first view, it shows daily utilization values for each site, broken down into pallet, depalletization, case, and fluid case categories. Since planned data only includes planned mono-ASIN pallets/units and case cases/units (which include both fluid cases and cases from mixed-ASIN pallets), we estimate fluid and mixed ASIN shares by using the historical data from the past four weeks to compute average proportions of mixed and fluid units and applying these shares to the planned volumes.

For each receive type, such as case or pallet, multiple subprocesses may be involved. For instance, case receive may include fluid unloading, PID conveyor, and PID scanning. To determine final utilization, we identify the most constrained subprocess, i.e. the bottleneck process, and report its utilization percentage as the summary value in the dashboard.

5.2 FC Maximum Processing Capacity Utilization

5.2.1 Introduction

The next step in the Inbound Maximum Processing Capacity Utilization project was focused on FCs. The goal remained the same: to give planners visibility into how much of the maximum processing capacity is expected to be used, but this time at FCs. However, the logic at FCs is more complex, as it involves more operational processes, such as stowing for inbound and picking for outbound.

Another layer of complexity comes from the type of inventory that arrives at FCs. As discussed earlier, FCs receive inventory not only from IXDs but also directly from sellers and vendors -

inventory that hasn't gone through IXD processing. On top of that, inventory arrives in various forms: pallets, cases, and totes. Totes typically contain units that have already been processed and sorted at IXDs, making them ready for stow without any additional handling.

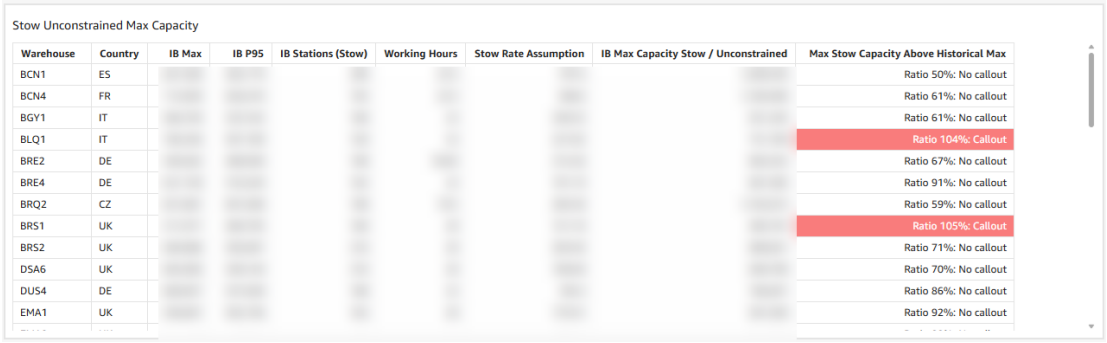
The final layer of complexity comes from FC-specific operational metrics that influence stowing and picking rates, both of which directly affect maximum and planned capacity. The key metrics here are bin fullness, drive utilization, and average cube per unit (ACU). Bin fullness, usually referred simply as fullness, measures how full the storage bins are. As bins become more full, stowing and picking rates slow down. Picking takes longer because it is harder to locate items, and stowing takes longer because there is less space to place new items. The same logic applies to ACU. Larger items (with higher cube per unit) reduce inbound capacity, as they take more time to process, particularly during decanting, when items are unpacked and placed onto the processing lines. Drive utilization measures how actively the robotic drives (which move bins to and from workstations) are being used. When drive utilization exceeds 95%, it typically leads to slower operations, as robots must wait longer to navigate the system. This in turn lowers both processing rates and overall capacity.

Despite these differences, the overall process followed the same structure as the IXD project: first, ACES developed a model to define the maximum processing capacities at FCs. Then we validated that model using historical peak 2024 data. Finally, we built a dashboard to show how much of that capacity is planned to be used.

5.2.2 Validation Analysis

The validation process focused on three main areas. First, we checked the unconstrained maximum stow capacity. Second, we looked at the maximum capacity based on the S&OP peak 2025 forecasted mix, which included the forecasted share of the TSI volume, and how much of that volume would be in totes. Third, we examined the residual IB capacity after accounting for the modeled OB maximum capacities, to make sure the overall balance made sense.

For the first part, unconstrained stow maximum capacity, we used a simple logic. We started by identifying the highest IB volume each FC site was able to achieve during Peak 2024. Then, we modeled the unconstrained IB maximum capacity by multiplying three factors assumed in the model: the number of IB stations (the total number of stow stations), the number of working hours per day, and the hourly stow rate. This gave us the estimated maximum capacity without any constraints (see Figure 14). Even though this was an unrealistic scenario, the goal of this analysis was to validate the model inputs. If a site's historical peak volume was higher than this modeled maximum, we flagged it for investigation. To validate the model inputs, we reached out to the FC POCs.


Stow Unconstrained Max Capacity

Warehouse	Country	IB Max	IB P95	IB Stations (Stow)	Working Hours	Stow Rate Assumption	IB Max Capacity Stow / Unconstrained	Max Stow Capacity Above Historical Max
BCN1	ES							Ratio 50%: No callout
BCN4	FR							Ratio 61%: No callout
BGY1	IT							Ratio 61%: No callout
BLQ1	IT							Ratio 104%: Callout
BRE2	DE							Ratio 67%: No callout
BRE4	DE							Ratio 91%: No callout
BRQ2	CZ							Ratio 59%: No callout
BRS1	UK							Ratio 105%: Callout
BRS2	UK							Ratio 71%: No callout
DSA6	UK							Ratio 70%: No callout
DUS4	DE							Ratio 86%: No callout
EMA1	UK							Ratio 92%: No callout

Figure 14: Validation of unconstrained IB maximum capacity. The table compares maximum peak 2024 inbound volumes achieved against modeled maximum stow capacity, calculated from the number of stow stations, working hours, and assumed stow rate. Sites where historical max exceeds modeled max by more than are flagged for further review.

In the second step, we compared the P95 historical inbound volume against the constrained maximum capacity (see Figure 15). Unlike the unconstrained values, constrained capacities

reflect more realistic assumptions, so using the P95 (instead of the max) provides a fairer comparison. This is important since capacity values heavily depend on factors such as product mix (pallet/case/tote volume split), and units per case, pallet, and tote.

Max Capacities basis S&OP Peak Mix

Warehouse	Country	IB Max	IB P95	Constrained Max IB Capacity	Tote Share Assumption	Constraint	Constraint Capacity	Callout
BCN1	ES					FLUID UNLOADING		Ratio 98%: No callout
BCN4	FR					FLUID UNLOADING		Ratio 118%: Deep Dive Required
BGY1	IT					DISTRIBUTION SORTER		Ratio 111%: Deep Dive Required
BLQ1	IT					CONVEYOR FROM DECANT/RECEIVE LINE		Ratio 87%: No callout
BRE2	DE					DISTRIBUTION SORTER		Ratio 99%: No callout
BRE4	DE					DISTRIBUTION SORTER		Ratio 87%: No callout
BRQ2	CZ					DISTRIBUTION SORTER		Ratio 93%: No callout
BRS1	UK					STOW		Ratio 88%: No callout
BRS2	UK					DECANT/RECEIVE LINE		Ratio 107%: Deep Dive Required
DSA6	UK					DISTRIBUTION SORTER		Ratio 92%: No callout
DUS4	DE					CONVEYOR FROM DECANT/RECEIVE LINE 1 & 2		Ratio 139%: Deep Dive Required
EMA1	UK					SPIRALS		Ratio 87%: No callout
EMA2	UK					DECANT/RECEIVE LINE		Ratio 80%: No callout
EMA4	UK					DECANT/RECEIVE LINE		Ratio 243%: Deep Dive Required

Figure 15: Validation of constrained IB maximum capacity based on S&OP peak 2025 mix. The table compares historical peak 2024 inbound volumes (P95) against modeled maximum IB capacity, which accounts for product mix and tote share assumptions (constrained). Sites where the ratio between the historical P95 volume and the modeled max exceeds 105% are flagged for deep dive, with the specific capacity constraint highlighted for each case.

We flagged any FC where the historical P95 exceeded 105% of the modeled max. For each flagged site, we identified the specific bottleneck process and its corresponding capacity. The next step was to validate whether the assumed values (i.e. model inputs) behind the bottleneck, such as the number of stations, hourly processing rate, working hours, assumed processes in the processing pipeline, were correct.

In the final validation step, we tested whether the modeled IB maximum capacity at a given OB volume matched actual performance during peak 2024 (see Figure 16). For each site, we began by identifying the modeled maximum OB capacity and calculating the corresponding modeled IB capacity at that OB level. Using the historical data, we then found days during peak 2024 when the actual OB volume was closest to the modeled OB maximum, keeping only those where OB was at least 90% of the modeled value. From this subset, we selected the day where the actual IB volume was closest to the modeled IB maximum. Comparing the modeled and actual IB volumes for these high-load days allowed us to assess the realism of the model's estimates. Sites where the historical IB volume fell outside the 85%–105% range of the modeled IB maximum were flagged for further review.

Residual IB Max Capacity at OB Max

Warehouse	Country	IB Max	IB P95	IB at Max OB Capacity	Max OB Capacity	Closest OB Achieved	Achieved OB / Max OB	Relative IB	Achieved IB / Max IB
BCN1	ES								
BCN4	FR								102%
BGY1	IT								100%
BLQ1	IT								102%
BRE2	DE								73%
BRE4	DE								80%
BRQ2	CZ								71%
BRS1	UK								100%
BRS2	UK								76%
DSA6	UK								89%
DUS4	DE								103%
EMA1	UK								105%

Figure 16: Validation of residual IB capacity at the modeled OB maximum. For each site, days during peak 2024 with OB volumes closest to the modeled OB maximum were identified, and the corresponding IB volumes were compared to the modeled IB maximum. Sites where the historical IB volume fell outside the 85%–105% range of the modeled IB maximum were flagged for review.

The reasoning behind this is simple: when outbound volume is high, inbound capacity is usually lower, since both compete for the same resources. For example, imagine a site that achieved 500K units outbound and 300K units inbound. If the modeled OB max is 450K and modeled IB max is 300K, we know the IB max is underestimated. That is because if we reduced OB to its modeled max (450K), IB should have gone up - exceeding the modeled IB max.

One important factor we had to consider was drive utilization, because it directly affects capacity and processing rates. Robots in FCs move pods that contain items to be stowed or picked. Drive utilization measures how much time these robots spend actively moving versus sitting idle. When drive utilization goes above 95%, it starts to impact performance. Robots take longer to reach their destinations because they need to wait for other robots to pass, which slows down both stowing and picking rates.

The challenge was to understand how drive utilization varied day by day. This was especially important during validation. For example, if we found that the historical inbound volume was significantly lower than the modeled max, e.g. below 85%, but drive utilization on those days was also low (below 85%), it meant there were no real constraints limiting inbound. In such cases, the modeled max was likely too high.

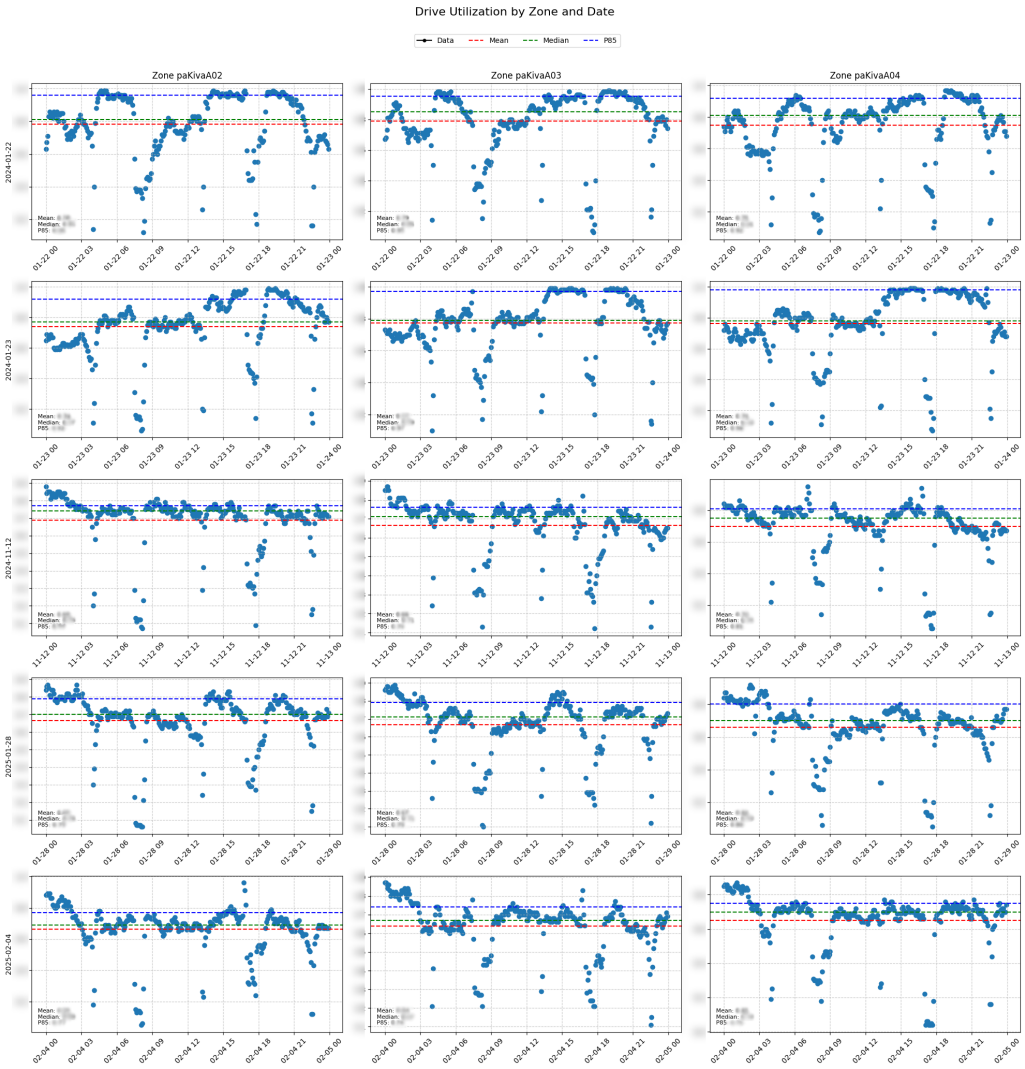


Figure 17: Drive utilization trends by zone (floor) and date. Each subplot shows the daily distribution of 5-minute interval drive utilization values for a specific zone. The dotted lines represent the daily mean (red), median (green), and P85 (blue) values. The consistent presence of extreme low and high values illustrates why max and median are not suitable daily statistics. P85 was selected as the most stable and representative value for daily drive utilization.

The drive utilization data we had was recorded at 5-minute intervals throughout the day (see Figure 17). To make this data usable for daily analysis, we needed to convert it into a single daily value. The key question was which metric to use - mean, median, max, P95, P85, etc.

We understood that using the max or P95 was not useful, since almost every site reached 100% drive utilization at some point during the day. These metrics would overstate the actual load. We then compared P85 with the median. However, the median was not a good fit either - each day had hours when utilization dropped to around 20%, which pulled the median down too much and did not reflect the overall picture.

In the end, we chose to use the P85 value. It effectively smoothed out the noise from extreme lows and highs, giving a more balanced and realistic view of daily drive utilization.

5.2.3 Utilization Dashboard

Following the same structure as in the IXD project, the planned capacity monitoring dashboard we developed for FC Maximum Processing Capacity Utilization included three main views. The first provided a general overview of how much of the total inbound capacity is planned to be used, also broken down into TSI and NVF capacities (see Figure 18). This view was similar to the ones used in the IXD dashboard. The second and third views were developed to explore different what-if scenarios and better understand why capacity may be under- or over-utilized at a given site.

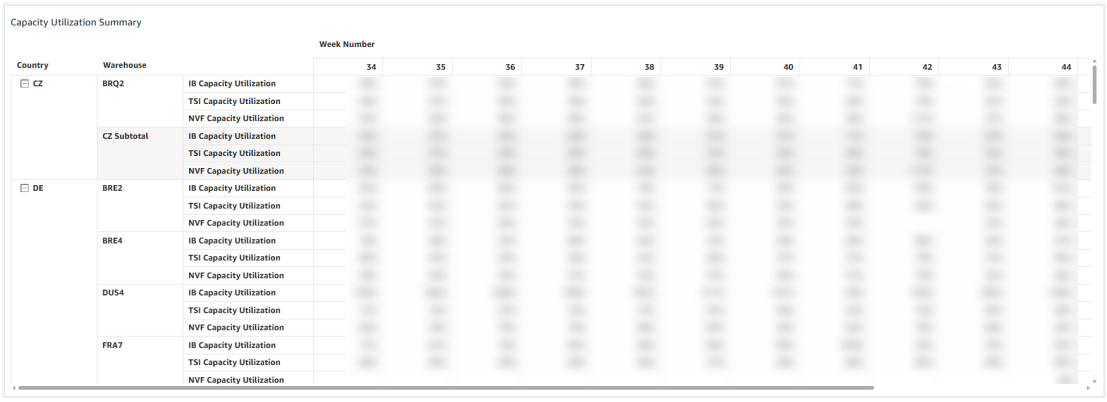


Figure 18: Capacity Utilization Summary view from the FC Maximum Processing Capacity Utilization dashboard, showing weekly planned utilization of inbound (IB), TSI, and NVF capacities across all FCs.

The second view allowed users to adjust the planned bin fullness, which directly impacts capacity (see Figure 19). As fullness increases, stowing rate slows down, reducing overall inbound capacity. The opposite happens when fullness decreases. Since stowing is the main part of the inbound receive process, any change in its efficiency directly affects how much inbound volume the site can handle. The goal of this view was to help users understand how sensitive capacity is to changes in fullness, and how close a site is to reaching its maximum capacity. This effect is known as degradation, and there are already models in place that quantify it. For example, increasing planned fullness from 90% to 95% may reduce the stowing rate by x units per hour, or by y percent. Note that all the calculations in the current planning models already include degradation factors, whether from fullness or ACU. These adjusted rates are referred to as degraded rates, or degraded TPH.

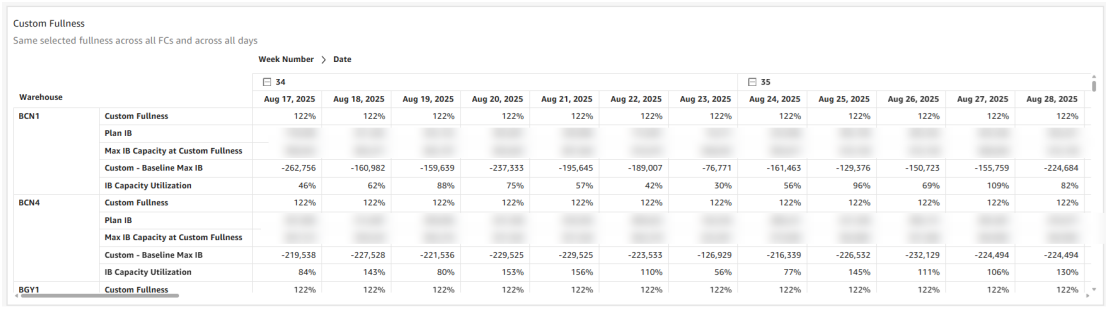


Figure 19: Custom fullness view from the FC Maximum Processing Capacity Utilization dashboard, showing the change of the maximum capacity when selecting different bin fullness values.

The final view focused on units per tote (UPT) (see Figure 20), which determines inbound processing efficiency, especially for stowing. Each inbound product type (also referred to as a "funnel") - such as NVF cases, NVF mixed ASIN pallets, NVF mono ASIN pallets, TSI totes, TSI cases, and various TSI pallet types - follows a specific process path before reaching the stowing stage. For example, all pallets are first unloaded at dock space. In contrast, totes and cases go through fluid unload and are then processed on a system called the dock sorter. Eventually, all funnels converge at the distribution sorter, which feeds into spiral conveyors and finally to the stowing area. Our task in this view was to model how the capacity of each of these processes shifts as UPT changes. The modeling logic followed a recursive structure. The total inbound capacity is essentially the sum of the capacities of all funnels. The capacity of each funnel is determined by the final process it passes through, which in turn depends on both its own output and the capacity of the processes before it.

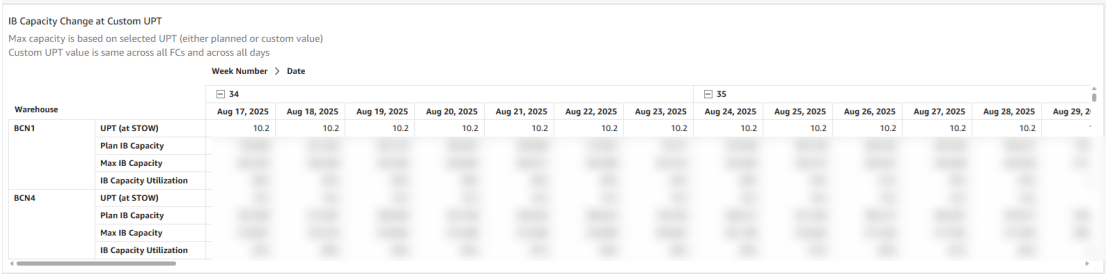


Figure 20: Custom UPT view from the FC Maximum Processing Capacity Utilization dashboard, illustrating how changes in UPT affect maximum capacity.

To estimate this, we used recent and planned data to approximate how much volume from each funnel type passes through each process. For example, we estimated how many NVF case units that were unloaded and toted would pass through the distribution sorter. The output capacity for a given funnel-process combination, such as NVF cases at the distribution sorter, was then determined by comparing the process's total available capacity with the allocated volume for that funnel. If the distribution sorter's maximum capacity was lower than the estimated total volume flowing through it, we calculated the funnel's output as the minimum between the output of the previous process in the chain (which varies by funnel) and the sorter's max capacity, scaled by the funnel's estimated volume share. Changing the UPT would increase or decrease the outputs of these processes if the process is a bottleneck, that is, if its maximum capacity is lower than the allocated volume. This change would then affect the output of the stowing step for that funnel, ultimately increasing or decreasing the overall inbound capacity.

5.3 Results

The work carried out addressed the three objectives defined at the start of the Inbound Maximum Processing Capacity projects.

1. Objective 1: Model validation. Using peak 2024 historical data, we identified and corrected discrepancies in model inputs, aligned static parameters such as workstation counts and throughput rates with actual site configurations, and ensured full input traceability. These improvements increased model accuracy and reliability, leading to approval for operational use by other teams.

2. Objective 2: Monitoring and scenario analysis. We developed interactive dashboards that compared planned volumes to modeled maximum capacities for both IXDs and FCs. Scenario analysis features allowed planners to adjust key parameters, such as pallet share, bin fullness, and units per tote, to assess the impact of different planning decisions.

3. Objective 3: Laying the foundation for optimization. While optimization was not implemented during the internship, the validated models and dashboards now serve as a basis for building algorithms that can balance volumes across the EU network.

The models and dashboards are scheduled for their first large-scale deployment during peak 2025, when demand is at its highest. Their use is expected to improve decision-making, reduce the risk of over- and under-utilization, and provide a more balanced network load during the most critical period of the year.

6 Discussion

6.1 Limitations and Challenges

One of the initial challenges was the steep learning curve in understanding the business model and operational structure of Amazon's supply chain. It took time to grasp how different planning processes, systems, and teams interact, which delayed the start of the analytical work.

Another limitation was the restricted time available for the Optimal Backlog Range Identification project. As a side initiative, it progressed only when primary project tasks allowed, which limited the depth of analysis that could be completed within the internship period.

For the Inbound Maximum Processing Capacity projects, a major challenge was the dependency on the models developed by another team. Model changes during the project required repeated adjustments to data pipelines and calculations, impacting timelines. Although this created rework, it also allowed us to identify and fix early issues in model assumptions, which strengthened the final version.

Data availability and reliability were another challenge. Even though there were many data sources, it was difficult to find datasets that were both reliable and suitable for our needs. Understanding the features and finding experts who could explain them took a lot of effort. In most cases, commonly used datasets lacked the detail we needed, so we had to look for less-known sources and check if they were accurate.

6.2 Areas for Improvement

A key improvement would be better project sequencing. Starting work before the model was finalized helped us find issues early, such as sortation process mismatches, which improved the final version. However, it also caused a lot of rework as the model and its inputs changed. In hindsight, it would have been more efficient to focus on another project, like the backlog project, until the model was stable, and then move to the capacity utilization work.

Another lesson learned concerns work delivery style. My approach often involved ensuring every detail of the logic was correct before sharing results. While thoroughness is valuable, this method sometimes delayed feedback. A more efficient strategy would be to share early, semi-finished versions of analyses, gather feedback iteratively, and then refine the logic for the final product. This approach would shorten feedback loops, improve alignment with stakeholders, and ultimately lead to a stronger final deliverable.

7 Conclusion

7.1 Summary

During the internship, I worked on two main initiatives: the Optimal Backlog Range Identification project and the Inbound Maximum Processing Capacity project. The backlog project was a side effort aimed at identifying the backlog range at which warehouses operate most efficiently. The method showed potential but would require further development and additional variables to improve robustness.

The Inbound Maximum Processing Capacity project was designed to optimize planned volumes across the entire EU Amazon network. The goal was to detect when a site was forecasted to exceed its maximum capacity and redistribute volumes to underutilized sites, thereby improving network balance. This work involved validating capacity models, developing interactive dashboards for monitoring and scenario analysis, and preparing for future optimization capabilities.

7.2 Future Directions

For the Optimal Backlog Range Identification project, future work should consider additional operational variables beyond the two features initially analyzed, as these alone were insufficient for stable results, particularly when historical data was limited.

For the Inbound Maximum Processing Capacity project, future improvements include integrating transportation cost considerations into the optimization process, as volume redistribution alone does not guarantee cost efficiency. Additionally, UPT modeling could be refined by focusing on initial inbound UPT for each funnel, rather than using aggregated or mixed values, and by analyzing the full sequence of processes for each funnel type, not just the final stages.

7.3 Final Thoughts

This internship provided an opportunity to apply my academic background in data science to large-scale operational challenges in one of the most complex supply chains in the world. While I did not work with advanced statistical models, machine learning, or deep learning, I learned the value of choosing the simplest effective solution, one that meets requirements, is interpretable, and can be operationalized quickly.

The experience also strengthened my skills in data validation, pipeline design, and visualization, and taught me the importance of clear communication between technical and business teams. Looking ahead, I plan to apply these lessons in future roles, combining technical expertise with a focus on operational impact.

References

[1] Amazon. 10 years of amazon robotics: How robots help sort packages, move product, and improve safety, 2022.

[2] Andy Jassy. Amazon ceo andy jassy's 2024 letter to shareholders, April 2025.

[3] Amazon Web Services. Amazon quicksight, n.d.

A Appendix

A.1 IXD Optimal Backlog Range Identification

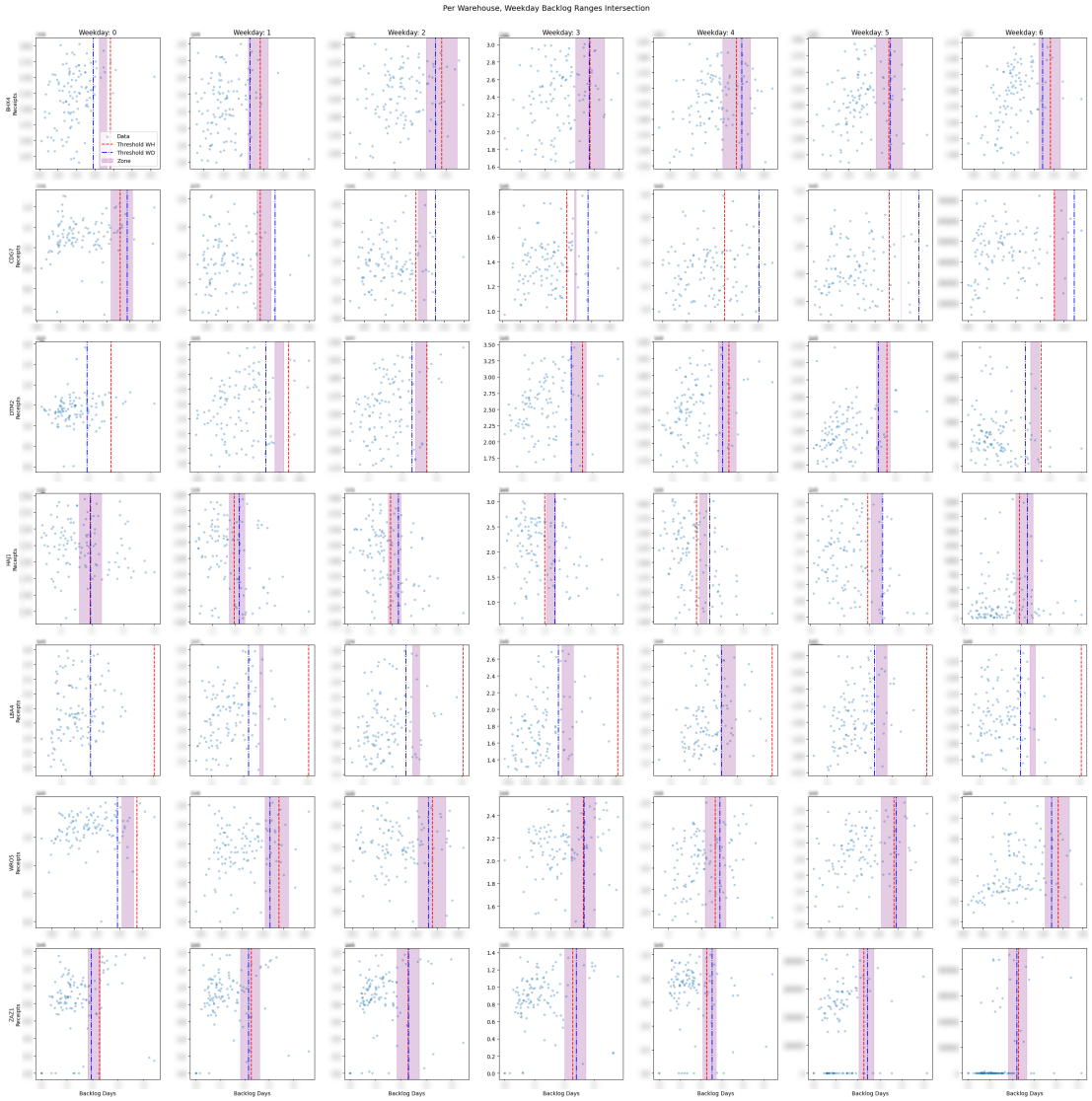


Figure 21: Intersection of optimal backlog ranges per warehouse and weekday. Each subplot represents a warehouse-weekday combination. The red dashed line indicates the threshold from the warehouse-level analysis, the blue dash-dot line represents the threshold from the weekday-level analysis, and the purple shaded region shows their intersection. If the intersection exists, it is interpreted as the recommended operational zone for that specific combination.

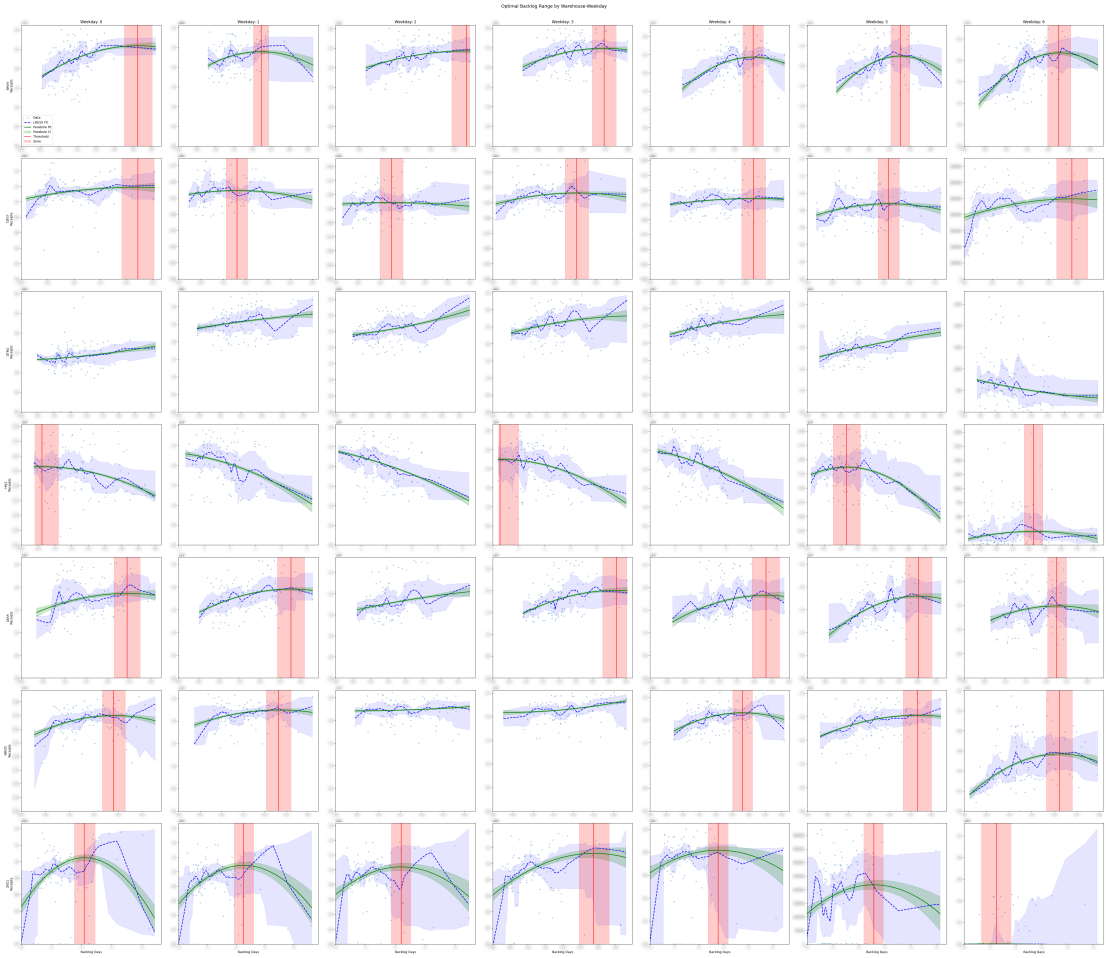


Figure 22: Optimal backlog ranges per warehouse-weekday combination. Each subplot shows LOESS smoothing (blue), a quadratic fit (green), and the resulting operational zone (red) for a specific pair. While some combinations yield desired results, others suffer from data sparsity or noise.