Facultat d'Informàtica de Barcelona

Master in Innovation and Research in Informatics

High Performance Computing

Statistical Modelling and Design of Experiments (SMDE)

HW 2

Name	Email
Qiuchi Chen	qiuchi.chen@estudiantat.upc.edu
Mihkel Tiks	mihkel.tiks@estudiantat.upc.edu
Mehmet Oguz Arslan	mehmet.oguz.arslan@estudiantat.upc.edu

Fall 24

Barcelona, June 1, 2025

Contents

1 Executive Summary

The objective of this assignment is to simulate the supply chain management of multiple walmart stores supplied by a central distribution center. The simulation will focus on inventory management at the individual stores, demand forecasting for various products, and the optimization of logistics for efficient product delivery. The goal is to reduce costs and loss of profits due to stockout.

The dataset that will be used is https://www.kaggle.com/datasets/mikhail1681/walmart-sales

2 System Description, Introduction

2.1 System Entity Definition

The system to be modeled consists of a central distribution center that supplies multiple individual Walmart retail stores.

- Distribution Center (DC): The DC serves as the central hub for receiving, storing, and distributing items to retail stores. Characteristics and functions of the DC are:
  - Processing store orders, including picking and packing.
  - Managing outbound shipments to individual stores.
- Retail Stores: Each store holds inventory to meet customer demand. Characteristics and functions of the retail stores are:
  - Receiving and stocking products from the DC.
  - Anticipating sales and avoiding stockouts.
  - Selling inventory to customers.
  - Has a certain capacity.

2.2 Core Process Modeling

- Product Flow: Products flow from the central DC to the individual stores. The simulation will model the time associated with order processing and transportation between the DC and the stores.

- Demand: Customer demand at the retail stores is the primary driver of the entire supply chain. Demand will be estimated based on the specific store, date, whether it is a holiday, the temperature in the region, fuel price, consumer price index and the unemployment rate.
- Logistics: The transportation of goods from the DC to the stores. The simulation will consider factors like transportation capacity, delivery schedules, and the cost associated with transportation. Optimization could involve minimizing transportation costs or delivery times.

3 Problem description

3.1 Core Goals

The problem we aim to address is the optimization of inventory management and logistics. This will be done taking into consideration the existence of multiple retail stores depending on one central distribution center. The aim will be to maximize profit through the following aspects:

- Minimize operating costs: including transportation costs (SH_01) and inventory holding costs (SH_02).
- Minimize out-of-stock losses: loss of sales opportunities due to insufficient inventory (SH_03)

3.2 Constraints

- Structural constraints:
  - Retail stores have limited inventory capacity (SS_04).
  - DC handling capacity and trucking capacity are fixed (SS_04).
- Data constraints:
  - Demand forecasts are based on historical data and external variables (SD_01).
  - Transport time distribution is subject to normal distribution assumptions (SD_03).

4 System Assumptions

4.1 Simplifying Hypotheses(SH):

- SH_01: Transportation Cost is calculated Per Unit and By Distance
- SH_02: DC to retail store truck speed and fuel consumption are constant
- SH_03: Trucks are loaded and unloaded instantly (?)
- SH_04: Items stored in Retail stores have fixed costs per day
- SH_05: Price of item is constant
- SH_06: Communication delay between DC and retail stores is negligible
- SH_07: Clients are served throughout the day evenly (with a constant interval with respect to the weekly sales)
- SH_08: The distribution center has unlimited supply

4.2 Systemic Structural Hypotheses (SS):

- SS_01: The distribution center goes over all the stores and checks whether they need to be resupplied once per day.
- SS_02: Retail stores either issue refill at certain inventory levels or according to a model
- SS_03: Orders are filled with a first come first serve policy (?)
- SS_04:

4.3 Systemic Data Hypotheses (SD):

- SD_01: Demand is estimated based on external factors:
- SD_02: Uniform Sales by Category, Using Granularity of Item:
- SD_03: Transportation time from DC to different stores follows normal distribution

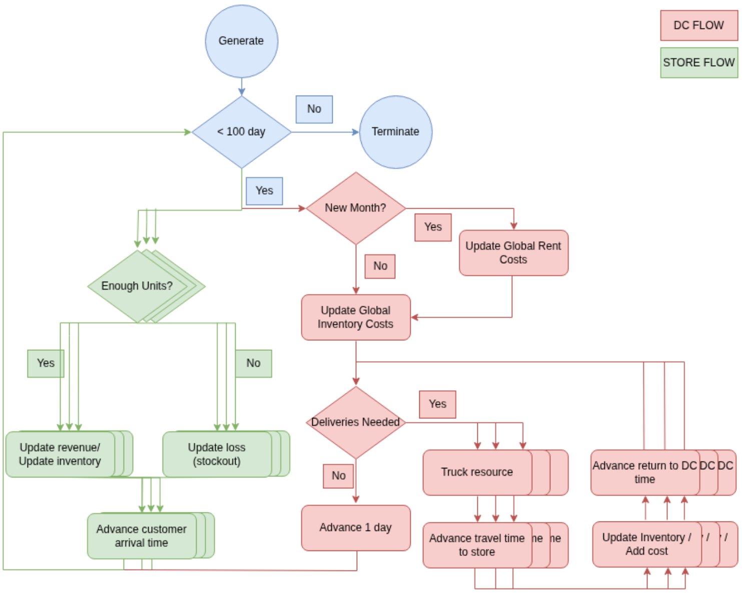

Figure 1: Flowchart describing both the actions of the distribution center and retail stores.

5 Model

5.1 Simulation Flow Chart

(need modify)

5.2 Demand forecasting model

- Objective: To predict customer demand at retail stores based on external factors as an input driver for supply chain simulation.
- Assumption mapping: Demand is related to stores, last week's sales, holidays, temperature, fuel prices, CPI, and unemployment rate (SD_01). The sales probability of the same category of goods is evenly distributed (SD_02).

5.2.1 Demand Prediction Using Multivariate Linear Regression

To model dynamic demand, we trained a multivariate linear regression model in R using historical data. The target variable is `Weekly_Sales`, and the predictors include:

- One-hot encoded `Store` index (Store1 to Store45)
- `Holiday_Flag`
- `Temperature`, `Fuel_Price`, `CPI`, `Unemployment`
- `Last_Week_Sales`

The regression model takes the form:

$$\widehat{y} = \beta_0 + \beta_{store_i} + \beta_H H + \beta_T T + \beta_F F + \beta_{CPI} C + \beta_U U + \beta_L L$$

Where:

- β_0 is the intercept
- β_{store_i} is the coefficient for the i-th store
- H = Holiday flag, T = Temperature, F = Fuel Price, C = CPI, U = Unemployment, L = Last week's sales

$$\widehat{y} \approx \text{Predicted Weekly Sales}$$

The predicted sales are then used in the simulation to compute expected hourly demand:

$$\lambda = \frac{\widehat{y}}{7 \times 24}$$

This λ is used to determine the rate of customer arrivals and to decide if a store is likely to run out of inventory within the next day. This enables the simulation to reflect demand-responsive logistics behavior.

5.2.2 Integration with Simulation

The model coefficients were extracted from R and manually embedded into the Python simulation. For each simulated store-day:

- Input features are retrieved from the historical dataset
- The regression formula is applied to calculate weekly demand

- This demand feeds into a decision rule to trigger delivery before a potential stockout

This hybrid modeling approach allows us to combine data-driven forecasting with process-level simulation, enhancing both realism and strategic insight.

5.3 Discrete Event Simulation Model

Simulate inventory flow, order processing and logistics transportation between DC and retail stores to verify the system structure hypothesis (SS) and simplification hypothesis (SH).

5.3.1 Solid Modeling

- 1. Distribution Center (DC):
    - Function: Receive store orders, pick and pack (processing time assumed to be fixed or normally distributed, SS_03), dispatch transportation (truck capacity constraints, SS_04).
    - Inventory Management: Process orders at fixed intervals (7 days), with limited inventory capacity.
- 2. Retail store:
- Inventory logic: When the inventory is below the safety line, replenishment is triggered (SS_{02}), the replenishment quantity is a fixed value (such as $safety\ line \cdot 2$), and the inventory capacity is limited (SS_{04}).
- Cost calculation: Daily inventory holding cost = $inventory \times fixed\ cost/day$ (SH_{02}), out_of_stock loss = $out\_of\_stock\ quantity \times fixed\ unit\ price$ (SH_{03}).
- 3. Logistics and transportation:
    - Time modeling: The transportation time from DC to store follows a normal distribution (Time $\sim N(\mu, \sigma^2)$), and the mean is positively correlated with the distance (SD_{03}).

  - Cost modeling: Transportation cost = transportation volume × distance × unit cost (SH_{01}), and the unit cost is fixed (SH_{03}, assuming the price is constant).

5.4 Event flow

- Demand event: Generate customer purchase requests based on forecasted demand, with a random fluctuation of ±10% (SD_{01}).
- Replenishment event: Place an order to the DC when the retail store inventory is $\leq$ the safety line (SS_{02}), and the DC processes the order at fixed intervals (SS_{03}).
- Transportation event: The truck is loaded with ordered goods, delivered to the store according to the transportation time model, and the inventory is updated.

5.5 Model output and verification

6 Coding

7 Definition of the Experimental Framework

A full factorial design was used for the experiments. Factors were chosen with the aim to represent different aspects of the supply chain to find out the most influential and efficient combinations.

7.1 Factors and Levels

For the simulations that were conducted, the following factors were considered:

- Storage conditions.
- Transportation conditions.
- Order policy.
- Demand volatility.

Factor	Level 1	Level 2	Level 3
A (Storage conditions)	Budget	-	Good
B (Transportation conditions)	Budget	-	Good
C (Demand Volatility)	Low	-	High
D (Order policy)	Static	Averaged	Linear Model

Table 1: Levels of factors involved in experiment

7.1.1 Storage conditions

For the 45 stores provided in the dataset, this factor represents different configurations that they receive throughout the simulations. The aspects that were modified were capacity and rent.

The capacity of a rental store would dictate how many items it could store at a time. Excess delivered items would be discarded. The values for this range from 50% (budget), 100% and 200% of the overall average weekly units sold. For monthly rents, the corresponding values are 500, 1000 and 2000.

7.1.2 Transportation conditions

Via the transportation condition factor a few parameters are determined - the amount of trucks (11, 22, 45 for the three levels), the speed of the trucks (50, 60, 70) and their maximum delivery quantity (20, 40, 60).

Maybe split these up into multiple factors? - speed, quantity...

7.2 Order Policies

Three different order policies were enacted and selected for the factorial design.

- Order when stockout would occur
- Order when below 50% of total capacity
- Order when free capacity larger than truck capacity

7.3 Design of Experiments

A reduced factorial design was used here to account for the otherwise large amount of combinations. The table of combinations can be found below. Each experiment was repeated x times.

7.4 Performance Metric

The performance metric that was measured in these tests was the total amount of profits. Using these combinations, better insight can be achieved about whether a business should prioritize for instance increasing their store inventory capacities, the trucks on the road or capacities of the trucks.

8 Model Validation

To ensure the correctness, credibility, and usability of our simulation model, we conduct several validation activities. These activities focus on verifying the data integrity, confirming the suitability of the experimental design, and assessing the ability of the model to mimic the behavior of the real-world Walmart supply chain. The following subsections detail each validation method used.

8.1 Data Validation

This step ensures that the datasets used in the simulation (e.g., sales, economic indicators, fuel prices) are accurate, relevant, and reflect the structure of the modeled system.

8.1.1 Data Distributions

The customer arrival process at each store is modeled using a Poisson process, with arrival intervals derived from historical weekly sales data. Sales data are divided by 100 (assuming 100 per product) to approximate the number of weekly customer arrivals. These arrival rates are assumed to be approximately normally distributed when aggregated, justifying the exponential interarrival times used in the simulation.

8.1.2 Data Sources

The data used in this simulation are obtained from a real-world Walmart dataset, commonly used in academic literature and available on platforms such as Kaggle. It includes weekly sales, fuel prices, CPI, and unemployment figures. These sources are considered reliable and representative of actual operational behavior.

8.1.3 Data Accuracy

We assume the dataset accurately reflects historical sales behavior and macroeconomic indicators. Although the data has been preprocessed (e.g., date formatting, weekly aggregation), we maintain data fidelity throughout the transformation process. No synthetic or artificially generated data were used.

8.2 Experimental Validation

This method ensures that our experimental procedures are designed to cover the relevant space of input parameters and that our simulation outputs are stable and statistically meaningful.

- **Experimental Coverage:** We conducted simulations across a variety of store storage types (small, medium, large), transport strategies (low-cost vs. high-capacity), and delivery decision policies (threshold-based vs. predictive model-based). This covers a wide range of realistic scenarios.

- **Stability and Repetitions:** For each configuration, the simulation was run over a sufficiently long period (e.g., 1000 days) to allow steady-state behavior to emerge. We also performed multiple runs with different random seeds to check for consistency and outlier-free behavior.

8.3 Operational Validation

Operational validation assesses the modelâs ability to realistically reproduce real-world system behavior, including demand patterns, stockout rates, and delivery frequency.

8.3.1 Regression-Based Demand Modeling

To closely mimic real-world demand variability, we integrated a multivariate linear regression model trained on historical data using R. The model includes store identity, economic indicators (e.g., fuel price, CPI), holiday flags, and previous week's sales as predictors. The model achieves meaningful predictive accuracy and captures temporal demand fluctuations at a weekly level.

8.3.2 Realism of Simulation Outputs

Key performance metrics (such as sales volume, stockouts, lost revenue, and transport costs) exhibit trends that are consistent with expected behavior under varying supply chain conditions. For example, predictive delivery models based on demand estimation showed reduced stockouts and better profit margins compared to static threshold-based policies.

8.3.3 Interpretation and Usefulness

The model allows practitioners to experiment with different delivery and storage strategies under realistic conditions. It facilitates cost-benefit trade-offs between inventory holding, transportation frequency, and service level. This ensures that the simulation provides not only a technically valid but also a decision-relevant representation of the Walmart retail distribution system.