الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic OF Algeria
وزارة التعليم العالي و البحث العلمي
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
جامعة الشهيد الشيخ العربي التبسي – تبسة
CHAHID CHEIKH LARBI TEBESSI UNIVESITY-TEBESSA
كلية العلوم و التكنولوجيا
Faculty of sciences and technology

DEPARTMENT OF ELECTRICAL ENGINEERING

Master's Thesis

Capstone defense to obtain Master's degree in:

Sector: Automatic
Specialty: Automatic and Industrial Information Technology

TOPIC

Traffic flow modelling with input/output data using B-Spline regression analysis

Presented on 08/06/2026 by the students :

AGOUNI Takoua
SASSI Asma

Panel of examiners :

President:	Mr. AMIEUR Toufik	Prof.	Echahid Cheikh Larbi Tebessi University
Supervisor:	Mr. DJABRI Riadh	MCB.	Echahid Cheikh Larbi Tebessi University
Examiner:	Mr. DJARI Abdelhamid	Prof.	Echahid Cheikh Larbi Tebessi University

2025/2026

Acknowledgment

First, we thank Almighty "ALLAH", for His kindness and giving.

We sincerely thank our supervisor,Dr.Djabri Riadh, for his valuable guidance, advice, and Continuous support

We also thank the members of the examination committee for accepting to evaluate this work.

Our deepest thanks go to our parents, families, and friends for their love, encouragement, and support

Dedication

We dedicate this modest work:

To our dear parents,

for their unconditional love, their sacrifices, their patience, and their endless support. Without them, nothing

would have been possible.

To our families, for their presence, encouragement, and trust throughout our academic journey.

To our friends, for the beautiful moments we shared, their moral support, and the positive energy that made this journey easier.

And to all the people who accompanied us, directly or indirectly, during our university life,

for their help, advice, and encouragement

ملخص

تُعدّ أنظمة المرور على الطرقات جزءًا أساسيًا من الحياة الحديثة، إلا أن الزيادة المستمرة في عدد المركبات أدّت إلى ظهور مشكلات مثل الازدحام والحوادث لذلك، أصبحت نمذجة تدفق حركة المرور أمرًا مهمًا لفهم وتحسين أنظمة النقل تتناول هذه المذكرة نمذجة تدفق حركة المرور باستعمال بيانات المدخلات و المخرجات ، مع إجراء مقارنة بين الطرق التقليدية وتقنية الانحدار باستخدام .B-Splines وقد تم استخدام بيانات حقيقية لحركة المرور جُمعت من الطرق السريعة في فرنسا ضمن الدراسة التطبيقية وأظهرت النتائج أن طريقة B-Spline Regression توفر نمذجة أكثر دقة ومرونة للسلوك غير الخطي لحركة المرور، مما يجعلها أداة فعالة للتنبؤ قصير المدى بحركة المرور

الكلمات المفتاحية : تدفق المرور ، نموذج ARX، نموذج ARMAX ، منحنيات B-splines ، التنبؤ بحركة المرور ، النمذجة غير الخطية

Abstract

Road traffic systems are essential in modern life, but the increasing number of vehicles has caused problems such as congestion and accidents. Therefore, traffic flow modelling is important for understanding and improving transportation systems.

This thesis studies traffic flow modelling using input-output data and compares traditional methods with B-Spline Regression. Real traffic data collected from highways in France were used in the practical study. The results showed that B-Spline regression gives more accurate and flexible modelling of nonlinear traffic behavior, making it an effective tool for short-term traffic prediction.

Keywords: Traffic Flow, ARX, ARMAX, B-Splines, Traffic Prediction, Nonlinear Modelling.

Résumé

Les systèmes de circulation routière sont essentiels dans la vie moderne, mais l'augmentation du nombre de véhicules a entraîné des problèmes tels que la congestion et les accidents. Par conséquent, la modélisation du flux de trafic est importante pour comprendre et améliorer les systèmes de transport.

Cette thèse étudie la modélisation du flux de trafic à l'aide de données entrée-sortie et compare les méthodes traditionnelles avec la régression B-spline. Des données réelles de trafic collectées sur des autoroutes en France ont été utilisées dans l'étude pratique. Les résultats ont montré que la régression B-spline offre une modélisation plus précise et plus flexible du comportement non linéaire du trafic, ce qui en fait un outil efficace pour la prédiction à court terme du trafic.

Mots-clés : Flux de trafic, ARX, ARMAX, B-Splines, Prévision du trafic, Modélisation non linéaire

List of symbols

q	Flow
k	Density
v_t	Time mean speed
v_s	Space mean speed
N	Number of veicles
n	Number of observation vehicles
d	Distance
t	Time
$y(t)$	System output
$u(t)$	Exogenous input
$e(t)$	White noise
θ	Parameter vector
$A(q^{-1})$	Autoregressive polynomial
$B(q^{-1})$	Input polynomial
$C(q^{-1})$	Moving-avrege polynomial
q^{-1}	Delay operator
$\hat{y}$	Predectted output
$\hat{\theta}$	Estimated parameter vector
H	Information matrix
Z	Instumental varaible matrix
X	Regression matrix
β_0	Regression intercept
β_1, β_2	Regression coefficients
$s(x)$	Spline function
$p_i(x)$	Polynomial segment of spline
t_i	Knot point
$c_{i,m}$	Polynomial coeffcients
$B_j(x)$	Spline basis function
$N_{i,p}(x)$	B-spline basis function
p	Degree of spline/B-spline
m	Number of knots(basis function)
u	Knot vector parameter
β_i	B-spline coefficients

List of abbreviations

AR	Autoregressive
MA	Moving average
ARMA	Autoregressive moving average
VAR	Vector autoregression
ARIMA	Autoregressive integrated Moving average
SARIMAX	Seasonal ARIMA with exogenous variables
ARX	Autoregressive with exogenous inputs
ARMAX	Autoregressive moving avrage with exogenous inputs
LS	Least squares
IV	Instrumental variables
MSE	Mean squared error
MRE	Mean relative error
SMS	Space mean speed
TMS	Time mean Speed
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion

List of figures

1.1 -Flow Density Relationship…….....6
1.2 - Comparison between Time Mean Speed and Space Mean Speed as a Function of Time…….7
1.3 - Illustration of a car-following situation showing the spacing between the leading and the following vehicle …….....9
1.4- Comparative analysis of macroscopic the speed-density models…….....16
1.5- Fundamental diagram showing capacity and flow-density relationship……16
1.6 - Microscopic level – Speed reduction propagation with 2-second reaction time delay …….17
2.1 - AR Model Structure…….21
2.2 - ARX Model Structure ……. ……...23
2.3- ARMAX Model Structure…….24
2.4 - Least square method – formula , definition , examples …….....25
2.5 - mbrenndoerfer.com ;Simple Linear Regression: Complete Guide with Formulas, Intuitive Explanations, Worked Examples, and Python Code"…….29
2.6 - data points that suggest quadratic regression would be a good fit……...30
2.7 - Comparative Analysis of LS vs. IV Estimation for ARMAX(2,1) Model……..34
2.8 - Prediction Results of ARMAX(3,1): Comparison Between LS and IV Methods……..34
2.9 - Impact of model type on model accury: linear,quadratic approach…….35
2.10 - Effect of Model Order on ARX Performance: A Study of ARX(2,1) and ARX(3,1)…….35
2.11 - Assessment of MAX Model Predictions Using Alternative Estimation Strategies……...35
3.1 - Mathematical Representation of a Spline Function S(x)…….39
3.2 - Linear Spline Interpolation……..42
3.3 - Quadratic spline……...42
3.4 - cubic spline…… …….43
3.5 - Traffic Flow Analysis: Quadratic vs. Cubic Spline with Lower Bound Constraints……...51
3.6 - Traffic Flow Modeling via Constrained Quadratic Programming……...51
4.1- Traffic flow modeling using B-spline fitting technique……55
4.2-B-spline regression and MRE Performance in traffiv flow testing……56
4.3- B-spline regression fit for real input/output traffic data flow……...57
4.4- MRE Evolution and B-Spline Model Tracking……58
4.5- Long-Term Traffic Flow B-Spline Modeling……..59
4.6- B-Spline Regression Performance and MRE……...60
4.7-Traffic folw modelling via B-splineregression……..61
4.8- B-Spline regression model performance during the testing phase……...63
4.9- Effect of parameter selection on B-spline regression for traffic flow modeling……..64

4.10-Reliability testing of the B-Spline model through MRE analysis and data fitting……..65
4.11-Dynamic estimation of traffic flow over time using cubic B-spline regression……..66
4.12- Validation of B-Spline regression for traffic flow prediction (MRE analysis)……...67

List of tables

1.1 - Comparison Between Time Mean Speed (TMS) and Space Mean Speed (SMS)……...7
1.2 - Automated Traffic Data Collection Technologies……13
1.3 - Observed speeds and travel times for a sample of five vehicles (Numerical Example)……14
1.4 - Computed traffic flow parameters based on the illustrative example……...15
2.1 - Traffic flow data……....32
2.2 - Estimation Results and Error Analysis of the ARX (2,1) Model……...32
2.3 - Comparative Performance of the ARX (3,1) Model Under LS and IV Methods……..33
2.4 - Error Metrics and Parameter Estimation for the MAX (2,1) Model……33
2.5 - Model Accuracy Assessment of the MAX (3,1) Structure……33
2.6 - Evaluation of the ARMAX (2,1) Model with Regression Benchmarks……34
2.7 - Predictive Performance Analysis of the ARMAX (3,1) Model……34
3.1 - Types of Continuity Conditions in Spline Functions……39
3.2 -Basic comparison of spline types……...43
3.3 - Classification of B-spline Types based on Knot Spacing……..48
3.4 - Comparative results of different Spline Regression models……..50
4.1- Number of Vehicles as a Function of Time and Velocity……53
4.2 - Training Phase Prediction Results……54
4.3 - Perfomance evaluation and testing……55
4.4 - Traffic Flow Testing and Validation Errors……..55
4.5 -Predictive Performance Evaluation During Training……56
4.6 - B-Spline Traffic Flow Model Validation Results……57
4.7 - Model Accuracy and Error Analysis in the Training Phase……..58
4.8 - Statistical metrics for the testing and validation phase……60
4.9 - Modeling results and statistical error……61
4.10 - Performance evaluation of the traffic flow model……..62
4.11 - Accuracy and error evaluation of the cubic regression model……63
4.12 - Evaluation metrics and prediction accuracy of the developed model over time intervals……..64
4.13 - Statistical efficiency evaluation of the model using MSE and MRE……..65
4.14 - Numerical results of the B-Spline regression model testing……...67

Summary

List of symbols i

List of abbreviations ii

List of figures iii

List of tables v

Summary vi

General Introduction 1

Chapter 1: Road traffic environment and conventional modelling approaches **4**

1.1. Introduction 5

1.2. Definitions and fundamental variables 5

1.2.1. Flow 5

1.2.2. Density 5

1.2.3. Speed 6

1.2.4. Fundamental relationships and traffic flow diagram 8

1.3. Traffic flow modelling with conventional methods 8

1.3.1. Microscopic model 8

1.3.1.1. car-following model 9

1.3.1.2. Lane-changing models 9

1.3.2. Macroscopic model 10

1.3.3. Limitations of conventional models 12

1.4. Collection and processing of traffic input–output data 12

1.4.1. Manual methods(field observation) : 12

1.4.2. Automatic methods (sensors) 13

1.5. Illustrative example 14

1.6. Conclusion 18

Chapter 2: Traffic flow modelling using parametric identification techniques **19**

2.1. Introduction 20

2.2. Time series modelling 20

2.2.1. Stationary and non-stationary time series modelling 20

2.2.1.1. Stationary 20

2.2.1.2. Non-Stationary 22

2.2.2. Models with exogenous inputs 22
2.3. Parametric identification 24
2.3.1. Least square method 24
2.3.2. Instrumental variables (IV) 27
2.4. Regression analysis 29
2.4.1. Linear regression 29
2.4.2. Quadratic regression 30
2.4.3. Model fitting and evaluation using training and testing data 31
2.4.4. Metrics error 31
2.5. Illustrative example 32
2.6. Conclusion 36
Chapter 3: Spline regression analysis **37**
3.1. Introduction 38
3.2. Mathematical foundations of spline function 38
3.2.1.Definition and properties 38
3.2.2. Key properties of Spline Function 40
3.2.3. Piecewise polynomial representation 40
3.2.4. Selection and placement of knots 41
3.3. Types of splines 41
3.3.1. Linear spline 41
3.3.2. Quadratic spline 42
3.3.3. Cubic spline 43
3.3.4. Comparative analysis of spline function 43
3.4. Spline regression model 44
3.5. B-spline regression 45
3.5.1. Historical development of B-Splines 45
3.5.2. Definition of B-spline 45
3.5.3. Key properties of B-splines 46
3.5.4. Types of B-splines based on knot vectors 47
3.5.5. B-spline using quadratic programming 48
3.6. Illustrative example 50
3.7. Conclusion 51
Chapter 4: Application on real input /output traffic data set **52**
4.1. Introduction 53
4.2. Data presentation 53
4.3.Simulation results 54
4.4. Conclusion 68
Conclusions and perspectives **69**
Références **70**

General Introduction

General introduction

Road traffic systems constitute a fundamental pillar of modern life, playing a vital role in facilitating the movement of people and goods while driving economic growth. However, the relentless increase in vehicle numbers, particularly within major urban centers, has birthed a series of complex challenges. These include chronic traffic congestion, escalating fuel consumption, environmental degradation, and a heightened risk of road accidents.

Consequently, developing robust methods to analyze and understand traffic behavior has become a priority for improving the management and efficiency of transportation networks. Traffic flow modeling seeks to describe and predict vehicle behavior through mathematical and statistical frameworks. These models primarily rely on three fundamental variables: flow (the number of vehicles), density, and speed. Various approaches have been established to study these systems, ranging from macroscopic models, which treat traffic as a continuous fluid-like system, to microscopic models, which focus on the intricate behavior of individual vehicles. While these models are foundational, they often depend on simplifying assumptions—most notably linearity. This often limits their capacity to capture the true complexity of traffic dynamics, particularly during the volatile transitions from free-flow to congested states.

To address these limitations, modern data-driven techniques have gained prominence, specifically the field of System Identification. These methods allow for the construction of models based directly on observed real-world data rather than relying solely on rigid theoretical formulas. Key parametric estimation techniques, such as Least Squares (LS) and Instrumental Variables (IV), are frequently employed alongside dynamic models like ARX and ARMAX. These models incorporate time dependencies and stochastic effects, providing a more faithful representation of evolving systems.

Nevertheless, real-world traffic systems are inherently nonlinear and complex, characterized by abrupt shifts in behavior that classical models may fail to capture. To provide greater flexibility, advanced techniques like Spline Regression, specifically B-Splines, have been introduced. Utilizing piecewise polynomial functions, B-Splines are highly effective at modeling nonlinear relationships. Their "local control" property is particularly advantageous, as it allows for adjustments in one segment of the model without compromising the integrity of the entire structure.

This dissertation focuses on traffic flow modeling using input–output data, providing a comparative analysis between conventional parametric identification methods and modern B- Spline approaches. The primary objective is to identify the most effective methodology for representing traffic dynamics, thereby enhancing prediction accuracy and model reliability.

To achieve this objective, the research is organized into four interconnected chapters:

-Chapter One establishes the general framework of traffic systems, defining the core variables (flow, density, and speed) and introducing the fundamental concepts of macroscopic and microscopic modeling.

-Chapter Two delves into system identification methods, covering parametric estimation (LS and IV) and dynamic models (ARX and ARMAX), emphasizing their ability to capture temporal and stochastic characteristics.

-Chapter Three is dedicated to Spline regression analysis, with a specific focus on B-Splines. It explores the mathematical properties and theoretical foundations that make this technique suitable for nonlinear modeling.

-Chapter Four presents the practical implementation and results. Real-world traffic data are utilized to apply the aforementioned models, followed by a comparative study of their performance and accuracy to determine the most suitable model for representing real traffic behavior.

Chapter 1

Road traffic environment and conventional modelling approaches

1.1. Introduction

Road traffic systems are essential in modern society. The increase in vehicle numbers has caused congestion and accidents due to the interaction between drivers, vehicles, and infrastructure. Therefore, traffic flow requires proper modelling to improve road performance. Traffic flow is described using three main variables: flow, density, and speed. These variables help explain traffic conditions and allow better evaluation of road performance.

Based on these variables, traffic is modelled using macroscopic, microscopic, and mesoscopic approaches. Macroscopic models focus on overall flow, microscopic on individual vehicles, and mesoscopic combines both. These models are useful but often rely on simplifying assumptions.

1.2. Definitions and fundamental variables

Traffic flow is defined as the study of vehicle movement on road networks. It focuses on the interaction between vehicles, drivers, and infrastructure. The main objective is to understand and design an optimal road network that ensures efficient traffic movement while minimizing congestion [1] In this context, traffic flow is characterized by three fundamental variables :

1.2.1. Flow

The number of vehicles passing a specific point on a highway during a given time period, usually measured in vehicles per hour (veh/h). [2]

$$q=\frac{3600N}{t_{measured}} \tag{1.1}$$

$t_{measured}$: time interval during which the vehicles were counted (in seconds)

N: number of vehicles

1.2.2. Density

can be defined as the number of vehicles occupying a given length of road at a given instant. Density is usually expressed as vehicles per kilometre

(veh/km). High densities indicate the vehicles are very close to each other, while low densities imply greater distances between vehicles [3]

$$K = \frac{n}{d} \tag{1.2}$$

k : traffic density [veh /unit distance]
n : number of vehicles on a part of roadway (section) at a specified time
d : length of the section

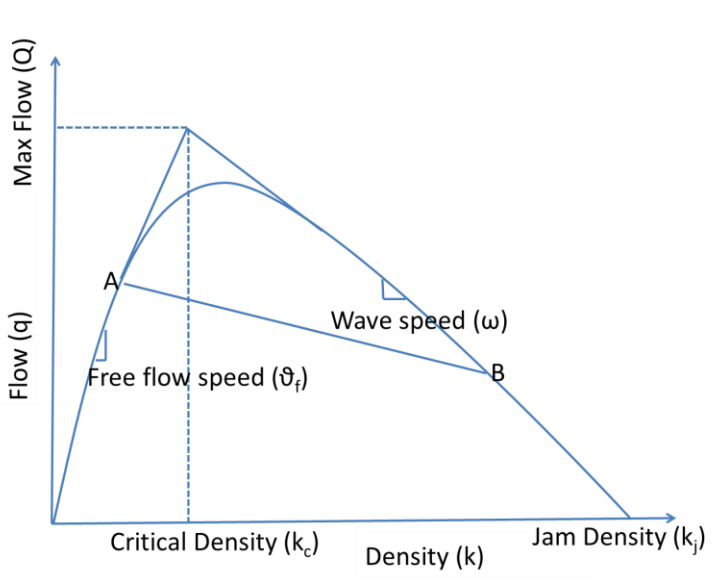


Figure 1.1-Flow density relationship

- ***Light density traffic:*** means there only a few vehicles on the road typically fewer than 400 vehicles per day . and that what making it easy to drive without congestion.
- ***Moderate density traffic :*** has more vehicles than light traffic ,between 400 and 1000 per day ,leading to some slowdowns but still manageable .
- ***High density traffic :*** has large number of vehicles closely packed together which is more than 1000 per day ,leading to congestion ,slower speeds , and potential delay [4]

1.2.3. Speed

defined as the average rate at which vehicles travel over a road segment, typically in km/h (kilometres per hour) or mph (miles per hour).

i.Time mean speed: is defined as the arithmetic average of the spot speeds recorded for all vehicles crossing a specific location during a predetermined time interval [2] This parameter is typically measured at a fixed point on the roadway using detection devices such as loop detectors embedded in the pavement. [5]

$$v_t = \frac{1}{N}\sum_{n=1}^{N} v_n = \frac{\sum_n (\frac{d}{t_n})}{N} \quad (1.3)$$

$v_n$is the spot speed of $n\ th$ vehicle

N is number of observations

d is distance traversed

t_n is $travel\ time\ for\ the\ n^{th}$ vehicles

ii.Space mean speed: represents the average speed of vehicles calculated over a specified road segment rather than at a single point. It is determined by observing vehicles as they travel along a defined length of roadway and computing the mean of their speeds within that segment at a given time. In practice, this measurement can be obtained using consecutive images or video recordings that track vehicle movements,

or through data collected from cameras and satellite-based systems. Because it accounts for the time vehicles spend traveling over a certain distance, space mean speed is generally considered more representative of actual traffic conditions than time mean speed. [2] [5]

$$v_s = \frac{N}{\sum_{n=1}^{N}\frac{1}{v_n}} = \frac{d}{\sum_n(\frac{t_n}{N})} = \frac{Nd}{\sum t_n} \tag{1.4}$$

Feature	Time mean speed(TMS)	Space mean speed (SMS)
Observation	Fixed point	Road segment
Calculation	Arithmetic Mean	Harmonic Mean
Bias	Favour faster vehicles	Representative of total flow

Table 1.1-Comparison between time mean speed (TMS) and space mean speed (SMS)

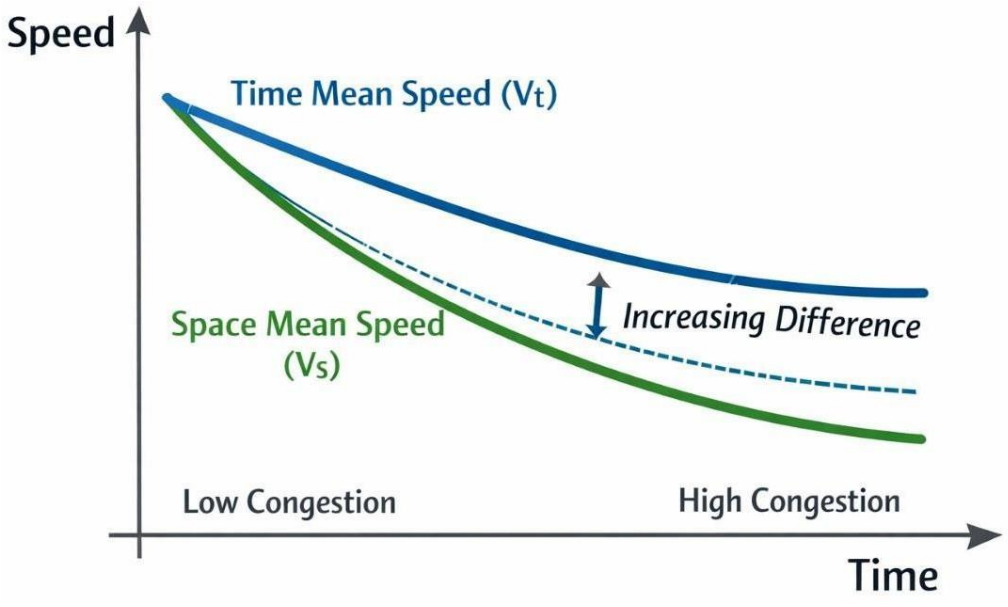


Figure 1.2-Comparison between time mean speed and space mean speed as a function of time

1.2.4. Fundamental relationships and traffic flow diagram

i.Fundamental relation : The formula below shows the relation between the variables:

$$q=k*u \tag{1.5}$$

with :

q : flow (veh /h)

k : density (veh/km) u : speed (km/h) [6]

ii.Fundamental diagram of traffic : it's a diagram that illustrates the relationship between the three primary traffic variable : flow, density and speed , it provides a graphical representation of how traffic behavior changes

In general:

At low density, speed remains high and flow increases.

At medium density, flow reaches a maximum value (capacity).

At high density, speed decreases significantly and flow reduces due to congestion.

This concept is central to traffic flow modelling and is often used to develop mathematical models. [7]

1.3. Traffic flow modelling with conventional methods

A traffic model is a mathematical representation used to describe and analyze the movement of vehicles on road networks. It helps researchers and engineers understand traffic flow behavior , predict congestion, and improve transportation system performance.

1.3.1. Microscopic model

Microscopic traffic flow models describe traffic dynamics by focusing on the behavior of individual vehicles . Every driver react to what happens around theme , such as the distance and speed of nearby vehicles or road elements like traffic lights and lanes. According to these conditions, drivers decide to accelerate, slow down, or change lanes, and the traffic flow results from these individual actions. characteristics of microscopic modelling methods were based on car-following model, lane-changing models. [8] [9]

1.3.1.1. Car-following model

Car-following models are microscopic traffic flow models that describe how a driver follows a leading vehicle. The following driver adjusts acceleration according to the spacing and the speeds of both vehicles. These interactions are usually represented using differential equations describing vehicle motion over time. [9]

The basic form of the model can be written as :

$$\dot{v_a} = F(v_\alpha(t), s_\alpha(t), v_{\alpha-1}(t)) \quad (1.6)$$

Where ;

F is car-following acceleratin vehicle

v_α is Speed of vehicle

s_α is distance between the vehicle and the leader

$v_{\alpha-1}$ is Speed of the Leading vehicle

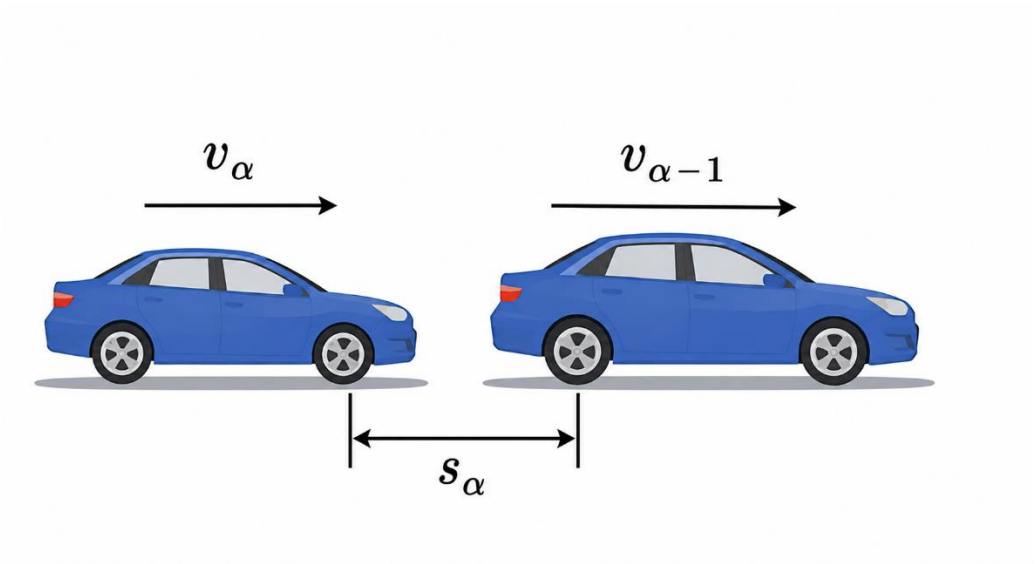


Figure 1.3-Illustration of a car-following situation showing the spacing between the leading and the following vehicle

1.3.1.2. Lane-changing models

Lane change models simulate how drivers decide to switch lanes by perceiving their environment, judging safety and speed benefits, and executing the manoeuver . They include discretionary (speed-driven) and mandatory (route-driven) changes, and can be represented using cognitive approaches like decision trees, neural networks, or information processing models. [9]

Basic Equation

i. Speed advantage (incentive criterion)

A driver considers changing lanes if the target lane offers a higher speed:

$$\Delta V = V_{target} - V_{current} \tag{1.7}$$

where:

V_{target} = speed in the target lane

$V_{current}$ = speed in the current lane

ii. Safety Condition (Gap Acceptance): The gaps in the target lane must be safe:

$$S_{front} \geq \ , S_{rear} \geq S_{safe} \tag{1.8}$$

where:

S_{front} = distance to the vehicle ahead

S_{rear} = distance to the vehicle behind

S_{safe} = minimum safe distance

iii. Acceleration Criterion : A lane change is accepted if the acceleration improvement is significant:

$$a_{new} - a_{old} > \Delta_{a_{th}} \tag{1.9}$$

where:

a_{new} = acceleration in the target lane

a_{old}= acceleration in the current lane

$\Delta_{a_{th}}$ = minimum acceleration advantage. [10]

1.3.2. Macroscopic model

Macroscopic traffic flow models represent traffic as a continuous flow characterized by aggregate variables such as flow, density, and mean speed, without explicitly describing individual vehicle behavior. These models are often used to analyze overall traffic patterns and can be represented through simple mathematical models, such as Greenshields' or Greenberg's. [9]

i. Greenshields model: Greenshields' Macroscopic Traffic Flow Model is one of the earliest and simplest models in macroscopic traffic flow theory. It describes the relationship between traffic speed and density, assuming that vehicle speed decreases linearly as traffic density increases. According to this model, vehicles travel at the free-flow speed when density is zero, while the speed becomes zero when the density reaches the jam density. This model helps explain how changes in density influence traffic flow behavior [11]

Equation:

$$v = v_f(1 - \frac{k}{k_j}) \tag{1.10}$$

Where ,

v is Mean Speed

v_f is free flow speed

k is traffic density

k_j is jam density

ii. Greenberg model: The Greenberg model is a macroscopic traffic flow model that expresses the relationship between traffic speed and density through a logarithmic function. It is mainly used to represent congested traffic conditions, where vehicle speed decreases as density increases. [11]

Equation:

$$v = v_c \ln(\frac{k_j}{k}) \tag{1.11}$$

$\boldsymbol{v_c}$ is critical speed

1.3.3. Limitations of conventional models

While the previously discussed macroscopic, microscopic models provide a solid theoretical foundation for understanding traffic dynamics, they possess several inherent limitations that affect their performance in real-world scenarios:

- Simplifying Assumptions: Conventional models often rely on simplifying assumptions that do not always hold true in complex traffic environments.

- Linearity Constraints: Models such as Greenshields' assume a strictly linear relationship between speed and density, which may fail to capture the nonlinear behavior observed in actual traffic flow.

- Reduced Flexibility: The reliance on predefined mathematical structures limits the flexibility of these models in accurately capturing complex and diverse real-world patterns.

- Congestion Modelling Gaps: While specific models like Greenberg's are designed for congested conditions, they may not bridge the gap effectively between theoretical frameworks and actual, fluctuating traffic behavior.

- Generalization Issues: Traditional methods may struggle to adapt to various road infrastructures and the diverse interactions between drivers and vehicles without significant recalibration. [12]

1.4. Collection and processing of traffic input–output data

Traffic data collection and processing leverage a range of techniques—from traditional manual counting to advanced automated systems—to capture essential metrics such as vehicle volume, speed, classification, and density for effective traffic analysis and management.

1.4.1. Manual methods(field observation) :

Manual traffic counting relies on human observers to collect qualitative data that automated systems often struggle to capture accurately, such as vehicle occupancy, pedestrian movement, and detailed vehicle classifications, particularly at intersections.

While this method is essential for gathering specific, nuanced details, it faces several key challenges compared to modern automated systems (like cameras and sensors):

_High cost: It requires significant human resources.

_Efficiency and accuracy: It is more prone to human error.

_Continuity: It lacks the capacity for continuous, long-term monitoring provided by more economically viable modern technologies. [13]

1.4.2. Automatic methods (sensors)

Method	Technology category	Principle of operation	Key advantages
Pneumatic tubes	Intrusive (surface)	Rubber tubes that detect air pressure changes when vehicle tires pass over them.	Low cost and ideal for temporary short-term studies.
Inductive loops	Intrusive (embedded)	Magnetic sensors buried in the pavement that detect the presence of metal.	Extremely high accuracy and works in all weather conditions.
Thermal imaging	Non-intrusive	Sensors that detect heat signatures emitted by vehicles.	High performance in total darkness, fog, or glare.
Radar technology	Non-intrusive	Uses radio wave reflections to measure vehicle speed and count.	Can monitor multiple lanes and is not affected by light conditions.
Artificial intelligence	Video-based (CV)	Uses computer vision to analyze video feeds and classify vehicle types.	Capable of complex classification (cars, trucks, bikes) and behavior analysis.

Table 1.2-Automated Traffic Data Collection Technologies [14]

1.5. Illustrative example

To illustrate the practical application of the fundamental traffic flow concepts introduced in this chapter, a two-step example is presented. The first part focuses on the computation of basic traffic variables using a simple dataset, while the second part extends the analysis to a realistic bottleneck scenario in order to demonstrate the interaction between macroscopic, mesoscopic, and microscopic traffic models.

Part I: Computation of Fundamental Traffic Parameters :

Consider a road segment with a length of d = 1 km. Five vehicles (N = 5) are considered , with the following speeds and travel times:

Vehicle	Speed (v_n) (km/h)	Travelled time (h)
Vehicle 1	120	0.00833
Vehicle 2	100	0.010000
Vehicle 3	80	0.01250
Vehicle 4	60	0.01667
Vehicle 5	40	0.02500

Table 1.3-Observed speeds and travel times for a sample of five vehicles (Numerical example)

- Time mean speed : using equation (1.3) $v_t = \frac{120+100+80+60+40}{5}$ =80 km/h
- Space mean speed : using equation (1.4) $\boldsymbol{v_s = \frac{5*1}{0.0725}}$ =68.97 km/h
- Density : using equation (1.2) k=5 veh/km
- Flow : using equation (1.1) q=344.85 veh/h

Parameter	Result
Time mean speed	80 km/h
Space mean speed	68.97 km/h
Density	5 veh /km
Flow	344.85 veh/h

Table 1.4 - Computed traffic flow parameters based on the illustrative example

These results show that the space mean speed is lower than the time mean speed, which confirms that slower vehicles have a greater influence on overall traffic conditions. The computed flow indicates a low-density traffic regime corresponding to free-flow conditions.

While this simple example provides a clear understanding of fundamental traffic variables, it remains limited to ideal conditions. In real-world traffic systems, disruptions such as lane reductions can significantly alter traffic behavior. Therefore, a more advanced scenario is required to capture these dynamics.

Part II: Bottleneck Dynamics and Multi-Scale Analysis :

To extend the basic numerical analysis presented in Part I, a more realistic traffic scenario is considered. A two-lane highway experiences a temporary lane closure due to maintenance, creating a bottleneck.

The upstream traffic conditions are defined as follows:

$$\text{Free-Flow speed} : v_f = 110 \text{ km/h}$$

$$\text{Density} : k_j = 120 \text{ veh/km}$$

- Analytical Solution:

The analysis is carried out at three levels: macroscopic, mesoscopic, and microscopic.

At the Macroscopic level by characterizing the global traffic stream. By applying the fundamental

parameters observed on the highway (v_f = 110 km/h, k_j = 120 veh/km), we compare the linear Greenshields model with the logarithmic Greenberg model. This comparison is crucial for understanding how speed declines as vehicles accumulate near the bottleneck entry As shows in figure 1.6. These models show that as density increases near the bottleneck, vehicle speed decreases.

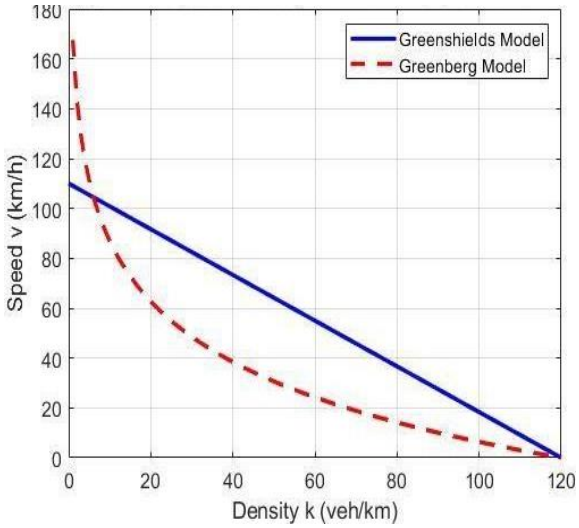


Figure 1.4- Comparative analysis of macroscopic speed-density models.

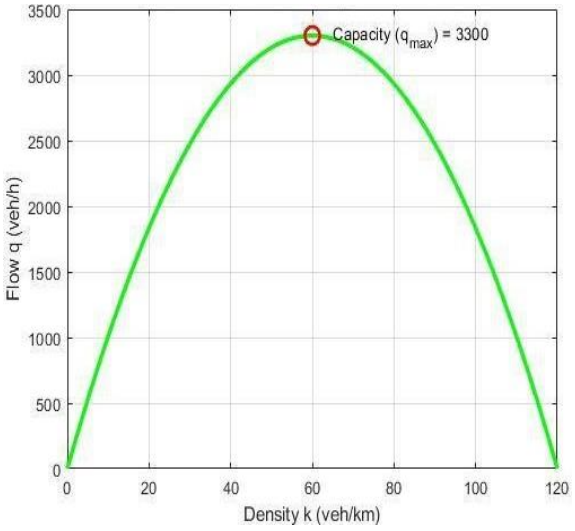


Figure 1.5- Fundamental diagram showing the capacity and flow-density relationship.

At the microscopic level, traffic behavior is explained using the car-following model. Due to the bottleneck:

-Vehicles reduce speed

-Inter-vehicle spacing decreases

-Drivers react with a certain delay (reaction time) These individual interactions lead to:

-formation of queues

-amplification of disturbances

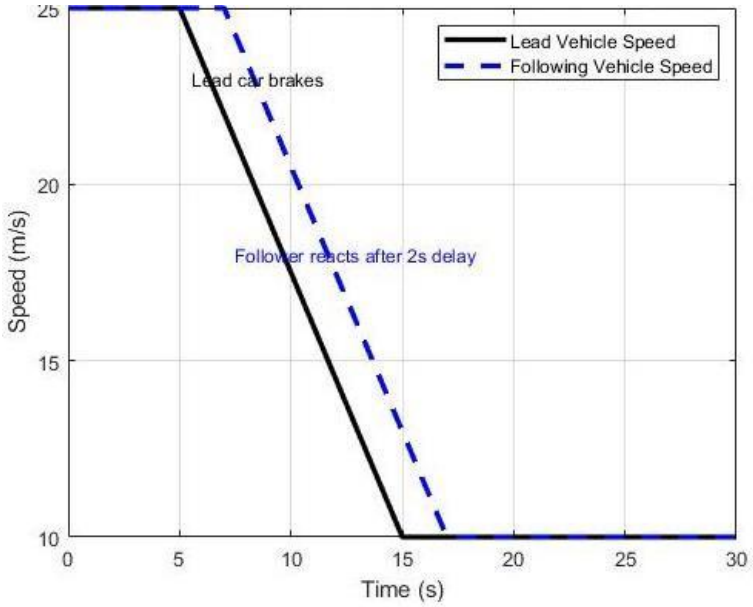


Figure 1.6- Microscopic level – speed reduction propagation with 2-second reaction time delay

This extended example demonstrates that traffic congestion in bottleneck scenarios cannot be fully explained using only basic variables. A multi-scale analysis is required to understand both the formation and propagation of congestion.

This limitation motivates the need to consider traffic flow regimes and phase transitions, which provide a more comprehensive description of congestion phenomena.

Traffic Flow Regimes and Phase Transitions :

The previous bottleneck scenario can be further interpreted using traffic flow regimes. According to the Three-Phase Traffic Theory proposed by Kerner, traffic can be classified into three distinct regimes:

-Free flow: Drivers can maintain their desired speeds as vehicle interactions are negligible.

-Synchronized flow: A congested state where vehicles move in a "synchronized" manner across lanes, and the flow rate remains relatively high despite lower speeds.

_Wide Moving Jam: A state of total congestion that propagates upstream as a shockwave, characterized by very low speeds and high density. The transition between these phases is often non-linear and exhibits "hysteresis" effects, meaning the path from free-flow to congestion is different from the path of recovery. Capturing these complex transitions requires flexible mathematical tools like B-Splines that can adapt to non-linear data trends. [15]

1.6. Conclusion

This chapter introduced the fundamental variables of the road traffic environment—flow, density, and speed—and reviewed conventional modelling approaches across macroscopic, microscopic, and mesoscopic scales. While these traditional methods provide a solid theoretical foundation, their reliance on predefined assumptions often limits their flexibility in capturing complex, real-world patterns.

To address these limitations, the next chapter explores traffic flow modelling using parametric identification techniques. These advanced methods offer a systematic approach to estimating model parameters directly from observed data, providing more accurate and adaptable models that effectively bridge the gap between theoretical frameworks and actual traffic behavior.

Chapter 2

Traffic flow modelling using parametric identification techniques

2.1. Introduction

Traffic flow modelling describes and predicts traffic behaviour using mathematical models based on time-dependent data. To estimate model parameters, identification techniques such as the Least Squares (LS) and Instrumental Variables (IV) methods are used. Additionally, regression analysis, including linear and quadratic regression, helps model relationships between traffic variables.
These techniques together provide a reliable framework for analysing, predicting, and improving traffic system performance.

2.2. Time series modelling

A time series model is a mathematical representation of a sequence of observations collected over time. These models are employed to analyze patterns in the data and to forecast future values. Time series modelling is widely applied in fields such as finance, economics, weather prediction, and epidemiology Time series models are commonly described through several key characteristics that reflect the nature of time- dependent data :

i.Autocorrelation : It is the similarity between observations based on the time lag separating them. It measures how closely current values relate to past values in a time series.

ii. Seasonality : Seasonality is a feature of time series data where a pattern repeats at regular intervals, such as seasonal changes.

iii. Stationarity : Stationarity is a key feature of time series. A series is stationary if its statistical properties remain constant over time. [16]

2.2.1. Stationary and non-stationary time series modelling

2.2.1.1. Stationary

means that the Statistical properties (the mean, variance, and autocovariance) are invariant over time. More formally, a time series x_t is strictly stationary if the joint probability distribution of any subset of the series remains unchanged when shifted in time. Several parametric models can be used to represent stationary time series [17]

i.Autoregression (AR) : An autoregressive model, noted AR(p), is used to forecast future values of a variable using its past values in a stochastic process, meaning it is not perfectly predictable.

The parameter p represents the number of lagged observations included in the prediction. If p = 1, the model uses only the previous observation in the series. In general, AR models rely on past values of the time series to estimate the current value.

$$A(q^{-1})y(t) = e(t) \tag{2.1}$$

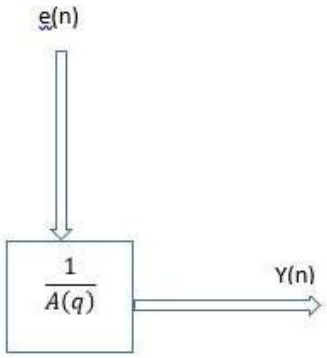


Figure 2.1 - AR model structure

ii. Moving average (MA) : A moving average model is a statistical method used to analyze and forecast time series data. Denoted as MA(q), it predicts future values using past forecast errors. The parameter q represents the number of previous errors included in the prediction. An MA(1) model uses one past error. In general, MA models rely on previous errors to estimate the current value. The general expression of an MA(q) model is given as: [18]

$$y(t) = C(q^{-1})\, e(t) \tag{2.2}$$

iii. Autoregressive moving average (ARMA) : ARMA models merge autoregressive components with moving average components. They are designed for stationary time series and are represented as ARMA(p, q). [18]

$$A(q^{-1})y(t) = C(q^{-1})e(t) \tag{2.3}$$

iv. Vector autoregression (VAR) : While (ARIMA) is suited for univariate tasks, vector autoregression (VAR) handles multivariate datasets.(VAR) models, such as (VARMA) and (VARMAX), can forecast multiple time series simultaneously [18]

2.2.1.2. Non-Stationary

a Non-S̄tationary Time Series is a sequence of data points where the statistical properties (such as mean, variance, and autocorrelation) change over time. In simpler terms, the "rules" governing the data shift as time passes, often due to external factors or underlying patterns. The main categories of non-stationary time series include [17]

i.Autoregressive integrated moving average (ARIMA) : ARIMA models are applied to non-stationary time series data, meaning data that exhibit trend or seasonality. Denoted as ARIMA (p, d, q), these models incorporate differencing to handle non-stationary series. [18]

*ii. Integration (I) :*Integrated models include differencing (d) to transform a time series into a stationary one. Differencing substitutes each value with the difference between the present value and a previous value, producing a new series that reflects the variation in the data. The parameter d specifies how many times the series is differenced [18]

iii.SARIMA with exogenous variables (SARIMAX) : More complex time series data involve variables beyond time. SARIMAX models include external variables to produce more detailed forecasts. [18]

2.2.2. Models with exogenous inputs

Unlike classical time series models such as (ARIMA) , and (ARMA) , which rely solely on past values of the output, ARX and ARMAX models incorporate exogenous inputs, making them more suitable for systems influenced by external variables.

i.ARX : ARX is a statistical model applied in time series analysis and in forecasting dynamic variables. ARX stands for “autoregressive model with exogenous inputs. [19]
The following equation shows the form of the (ARX) model.

$$A(q^{-1})y(t) = B(q^{-1})u(t) + e \tag{2.4}$$

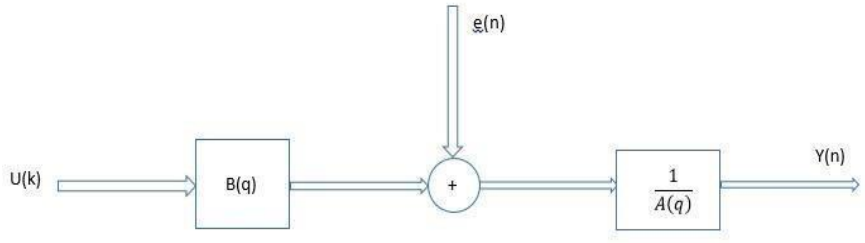


Figure 2.2 -ARX model structure

ii. ARMAX definition : A linear model employed in system identification and time series prediction. The notation ARMAX(p, q, b) denotes a model containing p autoregressive terms, q moving average terms, and b exogenous input terms. The final term represents a linear combination of the last b values of a known external time series. [20]

The following equation shows the form of the ARMAX model.

$$A(q^{-1})=\mathrm{B}(q^{-1})\mathrm{u}(t) + C(q^{-1})e(t) \tag{2.5}$$

Where

$$A(q^{-1}) = I - (A_1q^{-1} + \cdots + A_pq^{-1}) \tag{2.6}$$

$$B(q^{-1}) = B_1q^{-1} + B_2q^{-2} + \cdots + B_mq^{-p} \tag{2.7}$$

$$C(q^{-1}) = I + (C_1q^{-1} + \cdots + C_rq^{-r}) \tag{2.8}$$

$y(t)$ =System output

$u(t)$=Exogenous input

A, B, C=Polynomial in the delay q^{-1} operator

$e(t)$ = $white\ noise$

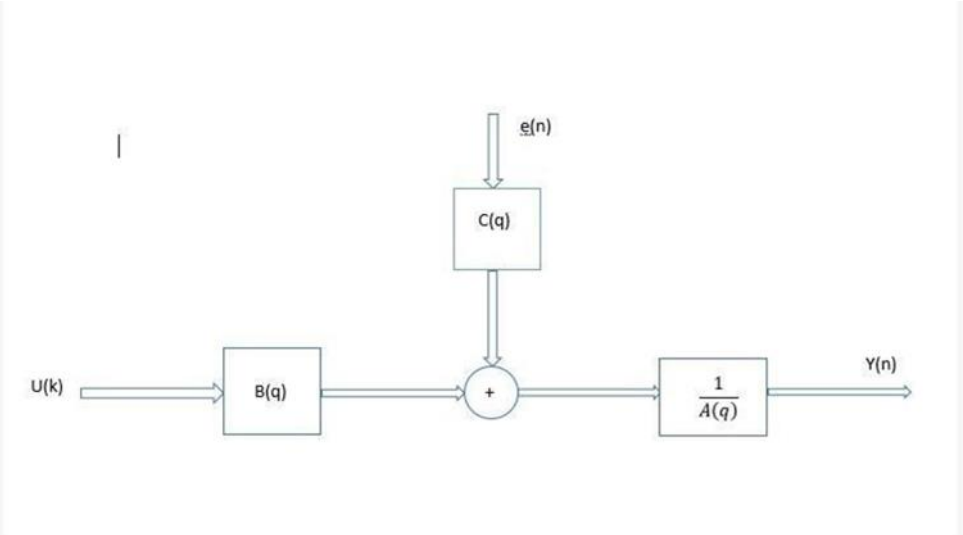


Figure 2.3 - ARMAX model structure

2.3. Parametric identification

Parametric identification is a technique used to estimate the unknown parameters of a mathematical model whose structure is predefined, using measured input–output data from a system. In this method, the model form is selected beforehand, and parameter estimation methods are employed to determine the parameter values that enable the model to accurately represent the system's behavior. [21]

Mathematically, a parametric model can be written as:

$$y(t) = f\ (u(t), \theta\ )\ e(t) \tag{2.9}$$

Where :

f ; model function

y(t) ; output

u(t) ; input

θ;vector of unknown parameters

e(t); white noise

2.3.1. Least square method

The Least Squares method is a widely used mathematical technique in data fitting, regression analysis, and predictive modelling. It identifies the best-fit line or curve, minimizing the sum of

squared differences between observed data points and predicted values, often expressed in the form

$$y = A_0\ \mathrm{u} + A_1 \qquad (2.10)$$

.

This curve is called the regression line, and the goal of the method is to minimize the total squared errors. [22]

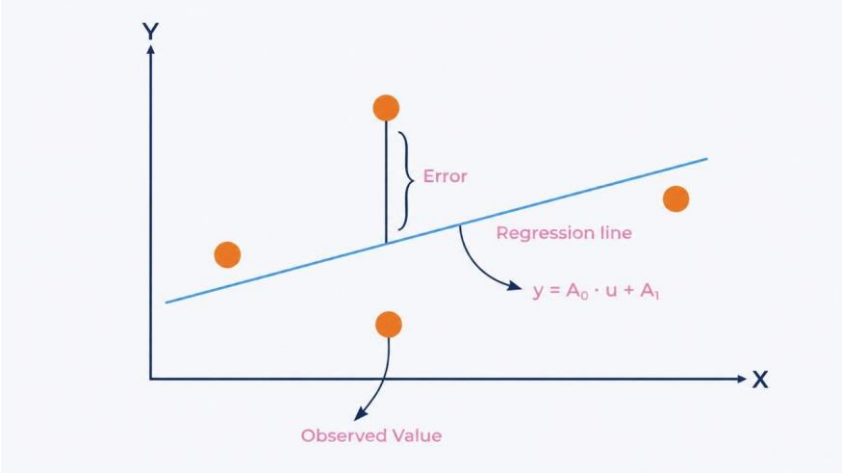


Figure 2.4 - Least square method - formula, definition, examples

Algorithm

Step 1 : Data collection

The first step involves gathering the experimental input-output dataset that represents the traffic dynamics over a specific period:

$$(\boldsymbol{u_i}, \boldsymbol{y_i})\ ; \boldsymbol{i = 1, 2, \ldots, m}$$

Step 2 : Model selection

$$\mathrm{y}(t) = b + a\ (t) + (t) \qquad (2.11)$$

We Select the model structure of type ARMAX

$$\mathrm{y(t)} = \mathrm{A_1 y_{t-1}} + \mathrm{B_1 u_{t-1}} + \mathrm{e(t)} \qquad (2.12)$$

where ,

$\mathrm{y_{t-1}}$ and $\mathrm{u_{t-1}}$ is past data

$\mathrm{e}(t)$ is white noise

Step 3: least squares optimization problem

To identify the most accurate parameters, we define an objective function based on the Least Squares method. The goal is to minimize the sum of squared residuals (the difference between actual and predicted outputs):

$$E_r = \mathrm{y(t)} - (\mathrm{A_1 y_{t-1}} + \mathrm{B_1 u_{t-1}}) \tag{2.13}$$

$$min_a \sum_{i=1}^{m} \cdot (\mathrm{y(t)} - \mathrm{A_1 y_{t-1}} - \mathrm{B_1 u_{t-1}}) \tag{2.14}$$

Step 4 :Parameter Estimation

To define the parameters vector of the chosen model (2.12):

$$\theta = \begin{pmatrix} A_1 \\ B_1 \end{pmatrix} \tag{2.15}$$

Step 5 :Matrix formulation

$$\widehat{y_t} = \widehat{A_1} y_{t-1} + \widehat{B_1} u_{t-1} \tag{2.16}$$

The matrix form of the estimation

$$\hat{y} = H.\theta \tag{2.17}$$

Where ,

H is Information Matrix of the input output

$\theta\hat{}$ is the estimate of parameter vector

$$H = \begin{pmatrix} y_1 & u_1 \\ \vdots & \vdots \\ y_{m-1} & u_{m-1} \end{pmatrix} \tag{2.18}$$

Step 6 :LS solution

The optimal parameter estimates are calculated using the standard Least Squares solution. This formula provides the values that statistically best fit the observed traffic data:

$$\theta = (H^T H)^{-1} H^T Y \tag{2.19}$$

[23]

2.3.2. Instrumental variables (IV)

Instrumental Variables (IV) is a statistical method commonly used in econometrics and causal inference to estimate causal relationships between variables when conventional regression techniques produce biased results. This problem usually occurs when an explanatory variable, called the independent variable, is correlated with the error term in the regression model, a situation known as endogeneity. The IV approach addresses this issue by using an additional variable, called an instrument, to obtain more reliable and consistent estimates. [24]

Algorithm

Step 1: Model specification

We consider the ARX model defined as:

$$y(t) = A_1 y_{t-1} + B_1 u_{t-1} + e(t) \tag{2.20}$$

This model can be written in a linear regression form:

$$y(t) = \varphi^T(t)\,\theta + e(t) \tag{2.21}$$

Where :

$$\theta = \begin{pmatrix} A_1 \\ B_1 \end{pmatrix} \tag{2.22}$$

$$\varphi(t) = \begin{pmatrix} y(t-1) \\ u(t-1) \end{pmatrix} \tag{2.23}$$

In matrix form

$$y = X\theta + e \tag{2.24}$$

Step 02

A suitable choice for ARX is:

$$z(t) = \begin{pmatrix} y(t-2) \\ u(t-2) \end{pmatrix} \tag{2.25}$$

The IV estimator is defined as:

$$\hat{\theta}_{IV} = (Z^T X)^{-1} Z^T y \tag{2.26}$$

Substituting $y = X\boldsymbol{\theta}$ into the IV estimator:

$$\hat{\theta} = (Z^T X)^{-1} Z^T (X\theta + e) \tag{2.27}$$

$$\hat{\theta} = \theta + (Z^T X)^{-1} Z^T e \tag{2.28}$$

Step 3 : Condition for unbiased

For the estimator to be unbiased:

$$E[\theta] = \theta \tag{2.29}$$

The following conditions must hold:

$$E[Z^T X] \neq 0 \quad \text{et} \quad E[Z^T e] = 0 \tag{2.30}$$

where

Z represents the matrix of instrumental variables.

X is the matrix of explanatory variables.

y is the dependent variable.

Step 4

The estimator is then computed as:

$$\hat{\theta} = (Z^T X)^{-1} Z^T y \tag{2.31}$$

[23]

2.4. Regression analysis

Regression analysis is a statistical technique for studying the relationship between a dependent variable (response) and one or more independent variables (predictors). It aids in prediction, forecasting, and understanding variable relationships.

2.4.1. Linear regression

Simple Linear Regression (SLR) is a statistical method used to model and quantify the relationship between a single independent variable and a continuous dependent variable. It is the simplest form of regression, which fits a straight line through the data points to minimize the differences between observed and predicted values, providing a clear representation of the linear relationship. [25]

$$Y = \beta 0 + \beta 1X + \varepsilon \tag{2.32}$$

Where:

Y is the dependent variable (DV or the outcome you're predicting).

X is the independent variable (IV or the predictor you're using).

$\beta 0$ is the intercept (predicted value of *Y* when *X* = 0).

$\beta 1$ is the slope (the average change in *Y* for a one-unit increase in *X*).

ε represents the random error or residual, indicating how individual observations deviate from the predicted values. [25]

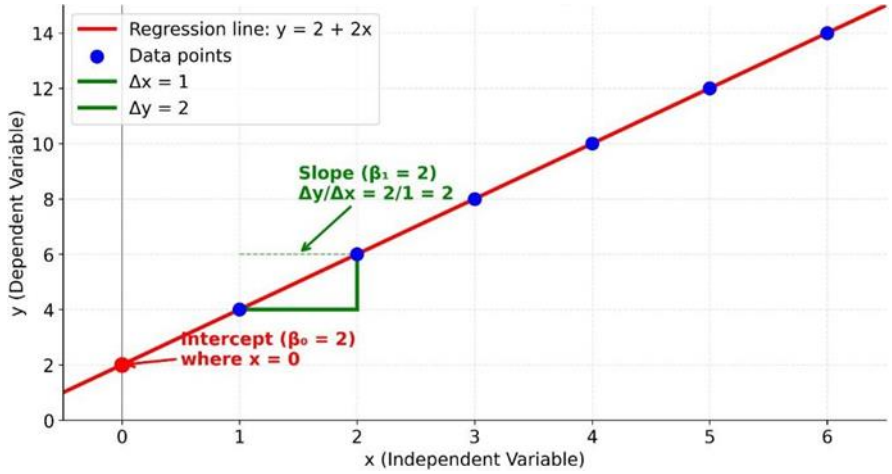


Figure 2.5-Simple linear regression: complete guide with formulas, intuitive explanations, worked examples, and python code"(mbrenndoerfer.com)

2.4.2. Quadratic regression

Quadratic regression is a form of polynomial regression that models the relationship between a continuous dependent variable and a single independent variable using a second-degree polynomial. This approach allows capturing nonlinear trends in the data that a simple linear model cannot represent. Although it models a curved relationship, quadratic regression is still considered a linear model because it remains linear in its parameters. [25]

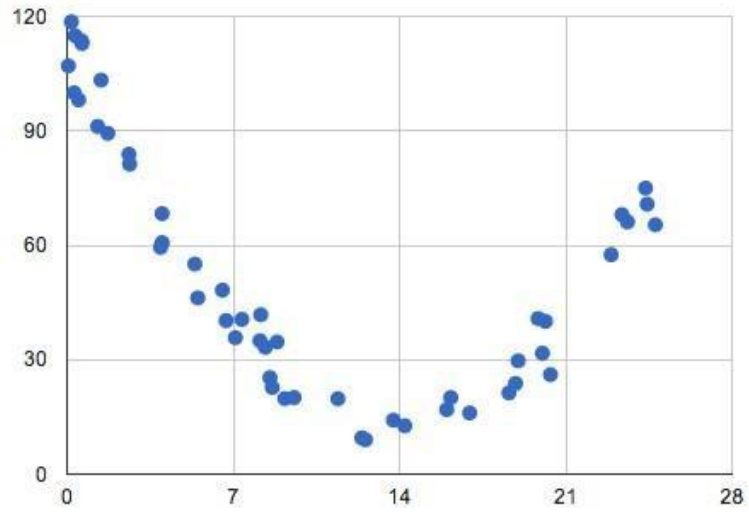


Figure 2.6- data points that suggest quadratic regression would be a good fit.

A typical quadratic regression model is written as:

$$Y = \beta 0 + \beta 1X + \beta 2X^2 + \varepsilon \quad (2.33)$$

Where:

Y is the outcome (dependent variable)

X is the predictor (independent variable)

β_0 is the intercept

β_1 is the coefficient for the linear term β_2 is the coefficient for the squared term

ε is the error term. [25]

2.4.3. Model fitting and evaluation using training and testing data

i.Line of Best Fit : A line of best fit is a line that most accurately represents the relationship between two variables on a scatterplot. It illustrates the overall trend in the data and is commonly used for prediction. While individual points may not lie exactly on the line, it is positioned to minimize the total distance between the points and the line, typically by minimizing the sum of squared residuals (the vertical differences between the observed points and the line). [25]

ii.Training Data : Machine learning algorithms learn by analyzing data contained in datasets. They detect patterns within the data, build an understanding of the relationships between variables, make decisions or predictions based on this information, and assess the accuracy of their results. The training data is the portion of the dataset used to train the machine learning algorithm and allow it to learn these patterns**.** [26]

iii.Testing Data : Test data refers to a collection of input values, conditions, and information used during the testing phase to verify the functionality of software applications. It is designed to represent real-world situations in order to check whether the software behaves correctly, manages errors properly, and maintains reliable performance under different conditions. Test data may include valid inputs for positive testing, invalid inputs for negative testing, boundary values, edge cases, and realistic user scenarios, ensuring thorough test coverage before the software is deployed in a production environment. [27]

2.4.4. Metrics error

Benchmarking metrics, testing and validation help optimize model performance, as they do in many other machine learning applications.

Time series modelling metrics include:

i.Mean squared error (MSE) : it is an evaluation metric that computes the average squared differences between actual and predicted values across all data points—the mean of the squared errors at each timestamp. [28]

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (2.34)$$

n is number of observations

i is the index of the sample

ii.Mean Relative Error : Mean Relative Error (MRE) is a statistical measure used to assess the accuracy of predictive models. It represents the average relative difference between the predicted values and the actual observed value [29]

$$MRE(\%) = \frac{100}{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (2.35)$$

2.5. Illustrative example

In this chapter, a simulated traffic flow dataset is used to illustrate time series models and regression analysis. The dataset contains traffic flow measurements taken every 15 minutes during morning hours, along with an external variable representing weather conditions.

Variables:

y = Traffic Flow (number of vehicles)

t= Time (hours)

Time	7.15	7.30	8.00	8.15	8.30	8.45	9.00	9.15	9.30	9.45	10.00
number of vehicles	117	94	164	160	186	143	160	182	183	156	167

Table 2.1 -Traffic flow data

- Model 1 : ARX (2.1)

$$y(t) = a_1\,(t-1) - a_2(t-2) + b_1u(t-1) + e_t \quad (2.36)$$
$$\theta = [a_1\ a_2\ b_1]$$

Estimation techniques	MSE	MRE (%)	Parameter estimation
LS	15.42	4.21	[0.252 0.518 0.8842]
IV	14.8	4.05	[0.231 0.4952 0.8520]
Linear regression	394.02	9.83	[112.4 6.3]
Quadratic regression	264.15	7.62	[89.2 15.6 -0.71]

Table 2.2-Estimation results and error analysis of the ARX (2,1) model

- Model 2 :ARX(3.1)

$$y(t)= a_1\, y_{t-2} - a_2 y_{t-2} + a_3 y_{t-3} + b_1 u_{t-1} + e_t \quad (2.37)$$

$$\theta = [a_1\ a_2\ \ a_3\ \ b_1]$$

Estimation techniques	MSE	MRE (%)	Parameter estimation
LS	13.20	3.85	[0.291 0.435 0.122 0.8504]
IV	12.65	3.52	[0.2785 0.412 0.1052 0.8240]
Linear regression	394.02	9.83	[112.4 6.3]
Quadratic regression	264.15	7.62	[89.2 15.6 -0.71]

Table 2.3-Comparative performance of the ARX (3,1) model under LS and IV methods

- Model 3 : MAX(2.1)

$$y((t)) = b_1(t-1) + e(t) + c_1 e(t-1) + c_2 e(t-2) \quad (2.38)$$

$$\theta = [\ b_1\ \ c_1\ \ c_2\ ]$$

Estimation techniques	MSE	MRE (%)	Parameter estimation
LS	128.35	6.92	[5.83 3.12 1.15]
IV	-	-	-
Linear regression	394.02	9.83	[112.4 6.3]
Quadratic regression	264.15	7.62	[89.2 15.6 -0.71]

Table 2.4 -Error metrics and parameter estimation for the MAX (2,1) model

- Model 4 :MAX(3.1)

$$\mathrm{y}(t) = b_1(t-1) + e(t) + c_1 e(t-1) + c_2 e(t-2) + c_3 e(t-3) \quad (2.39)$$

$$\theta = [b_1\ \ c_1\, c_2\, c_3]$$

Estimation techniques	MSE	MRE (%)	Parameter estimation
LS	110.20	5.85	[5.20 2.85 1.45 0.95]
IV	-	-	-
Linear regression	394.02	9.83	[112.4 6.3]
Quadratic regression	264.15	7.62	[89.2 15.6 -0.71]

Table 2.5-Model accuracy assessment of the MAX (3,1) structure

- Model 5 : ARMAX (2.1)

$$y(t) = a_1(t-1) + a_2y(t-2) + b_1u(t-1) + e(t) + c_1e(t-1) \quad (2.40)$$
$$\theta = [\,a_1\ b_1\ c_1]$$

Estimation techniques	MSE	MRE (%)	Parameter estimation
LS	13.92	3.90	[0.2450 0.5002 0.8150 0.1500]
IV	13.45	3.72	[0.2380 0.4850 0.7920 0.1420]
Linear regression	394.02	9.83	[112.4 6.3]
Quadratic regression	264.15	7.62	[89.2 15.6 -0.71]

Table 2.6-Evaluation of the ARMAX (2,1) model with regression benchmarks

- Model 6 : ARMAX (3.1)

$$y(t) = a_1(t-1) + a_2y(t-2) + a_3y(t-3) + b_1u(t-1) + e(t) + c_1e(t-1) \quad (2.41)$$
$$\theta = [a_1\ \ a_2\ \ a_3\ b_1\ c_1]$$

Estimation techniques	MSE	MRE (%)	Parameter estimation
LS	11.45	3.20	[0.260 0.420 0.095 4.920 0.183]
IV	11.12	3.05	[0.252 0.405 0.088 0.762 0.175]
Linear regression	394.02	9.83	[112.4 6.3]
Quadratic regression	264.15	7.62	[89.2 15.6 -0.71]

Table2.7-Predictive performance analysis of the ARMAX (3,1) model

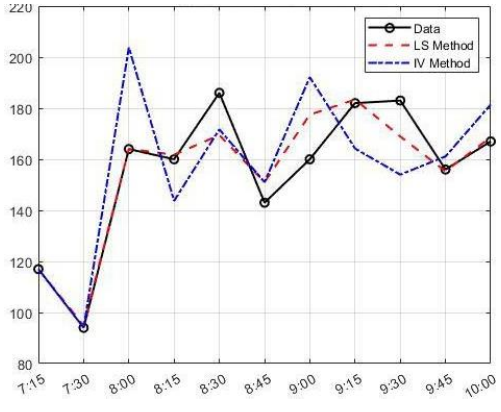


Figure 2.7 - Comparative analysis of LS vs. IV estimation for ARMAX(2,1) model

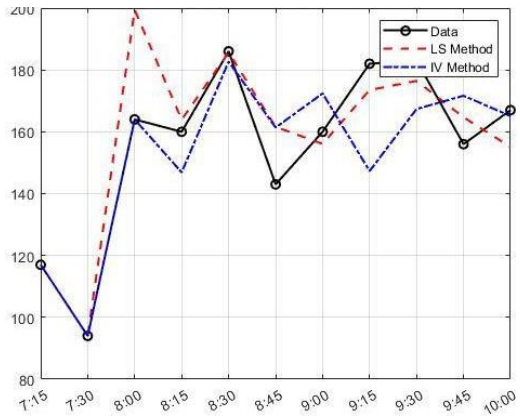


Figure 2.8 -Prediction results of ARMAX(3,1): comparison between LS and IV methods

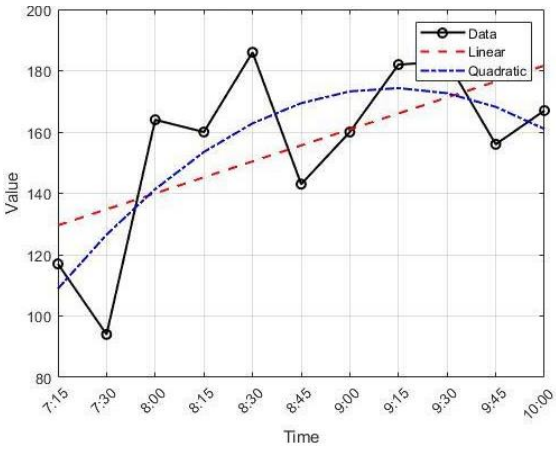


Figure 2.9-Impact of model type on model accury : linear,quadratic approach

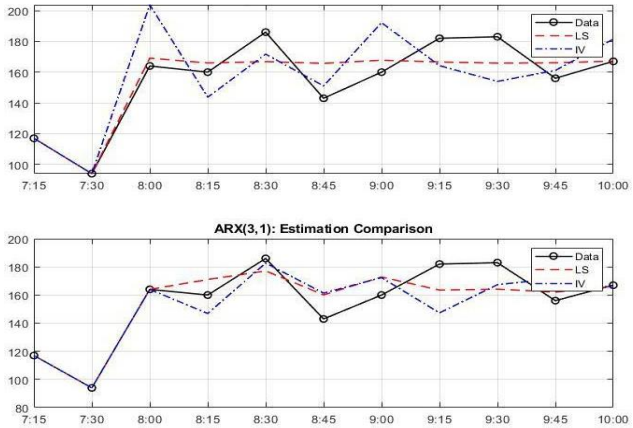


Figure 2.10-Effect of model order on ARX performance:
a study of ARX(2,1) and ARX(3,1)

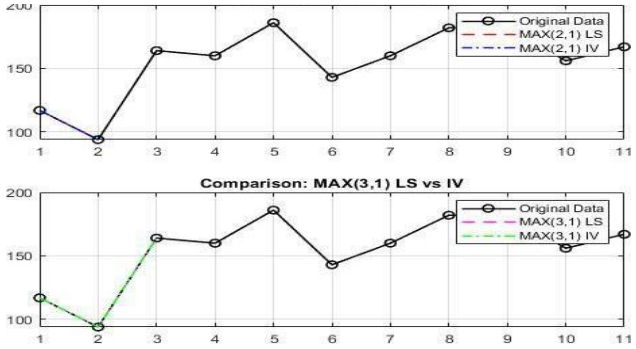


Figure 2.11 -Assessment of MAX Model predictions using alternative estimation strategies

The dataset exhibits significant temporal variability, indicating that traffic flow is inherently dynamic. The variation in vehicle counts confirms the presence of complex, time-dependent patterns. The inclusion of weather conditions as an exogenous variable enhances the explanatory power of the model.

Static modeling approaches are insufficient to capture the observed data behavior.

The ARMAX (3,1) model achieved the highest predictive accuracy among all tested models. Its performance reflects the importance of combining autoregressive, moving average, and exogenous components.

The ARX (3,1) model also demonstrated strong predictive capability with slightly lower accuracy.

The results confirm that incorporating temporal dependencies significantly improves model performance.

Linear regression failed to adequately represent the nonlinear nature of traffic flow. Quadratic regression improved model fit but remained insufficient due to the absence of temporal structure.

These findings highlight the limitations of static regression models in dynamic systems. The Instrumental Variables (IV) method consistently outperformed the Least Squares (LS) approach.

The improvement in accuracy demonstrates the robustness of IV in dynamic modeling contexts.

The superiority of IV is attributed to its ability to address endogeneity issues.

LS estimation is prone to bias when explanatory variables are correlated with the error term. time-series approaches..

The ARMAX (3,1) model is identified as the most suitable model for traffic flow prediction. The IV estimation method is recommended for obtaining consistent and unbiased parameter estimates.

2.6. Conclusion

This chapter highlighted how parametric methods, including LS and IV for time series, along with linear and quadratic regression, estimate traffic flow parameters and capture basic traffic dynamics. Yet, real traffic often shows nonlinear patterns that these models cannot fully address. The next chapter introduces spline regression, using piecewise functions to model abrupt changes, offering more precise forecasting and system optimization.

Chapter 3

Spline regression analysis

3.1. Introduction

This chapter presents the fundamental principles of spline-based modelling, including linear, quadratic, and cubic splines, followed by an introduction to B-spline regression as a robust and flexible approach for handling complex data structures.Spline functions provide a flexible solution by using piecewise polynomials connected at knots with smoothness conditions. This chapter introduces linear, quadratic, and cubic splines, and presents B-spline regression as an efficient and stable approach for modelling complex data.

3.2. Mathematical foundations of spline function

3.2.1. Definition and properties

A spline is a piecewise polynomial function defined over a partitioned interval, where each segment is a low-degree polynomial, and continuity and smoothness are enforced at the junction points called knots. [30]

$$s(x) = \begin{cases} p_1(x) & \begin{cases} \text{if } x \le t_1 \\ \text{if } t_1 < x \le t_2 \\ \vdots \\ \text{if } x > t_d \end{cases} \\ p_2(x) \\ \vdots \\ p_{d+1}(x) \end{cases} \tag{3.1}$$

$s(x)$: This is our spline function

d: This represents the degree of the spline. t1,t2,…,t_d: These are called knots or knot points.

p_i (x): polynomial of degree d or less

$$\boldsymbol{P_i(x) = a_i x^1 + b_i + c_i x^2} \tag{3.2}$$

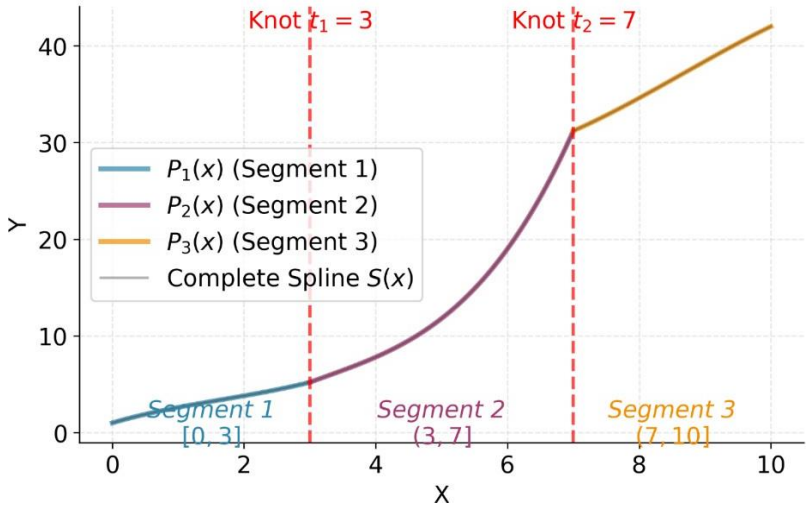


Figure 3.1- Mathematical representation of a spline function S(x)
(Spline regression: complete guide to non-linear modeling with mathematical foundations & python Implementation michael brenndoerfer June 18, 2025)

Splines are widely used because traditional high-degree polynomial models suffer from issues such as oscillations (Runge's phenomenon) and lack of local control.

i.Continuity Condition : Continuity conditions can be summarized as constraints that ensure smooth connections between consecutive polynomial pieces at each knot. [31]

Order	Condition	Meaning
C^0	$P_i(x_i) = P_{i+1}(x_i)$	No jump (Function is continuous)
C^1	$P_i(x_i) = P_{i+1}(x_i)$	No sharp corner (Same slope)
C^2	$P_i(x_i) = P_{i+1}(x_i)$	Same curvature (smooth bending)

Table 3.1- Types of continuity conditions in spline functions

3.2.2. Key properties of Spline Function

Spline functions possess several important properties that make them highly effective for modelling complex systems :

i. *Local control* : B-spline ensure local control of the curve shape, such that any modification of a control point ci affects only a limited region of the curve in the vicinity of that point.

*ii. Numerical Stability :*A B-spline curve always lies within the convex hull of its control points. Moreover, each segment of the curve is contained within the convex hull defined by the subset of control points that influence that segment. Specifically, for $y = [t_i, t_{i+1})$, the point (y)belongs to the convex hull of C_{i-p}, C_{i-p+1}, … . , C_i

*iii. Smoothness :*A B-spline curve of degree $p \geq 1$ has continuity C^{p-1}. For instance, a cubic B-spline exhibits C^2 continuity. However, increasing the knot multiplicity r reduces the continuity at the corresponding knot to C^{p-r}. The smoothness of the curve is therefore determined by the knot multiplicities, while within each interval $[t_i, t_{i+1})$, the curve remains infinitely differentiable. [32]

3.2.3. Piecewise polynomial representation

In practice, a spline is formed by using different polynomials on each interval of the domain, each defined around its own knot.

This local formulation can be written as:

$$\boldsymbol{P_i(x)} = \sum_{m=0}^{d} c_{i,m}(x - x_i)^m \qquad (\mathbf{3.3})$$

Where ,

$\boldsymbol{c_i}$, are coefficients specific to each interval,

$\boldsymbol{x_i}$ is the left knot of the interval,

$\boldsymbol{d}$ is the polynomial degree.

The full spline is then obtained by combining all local polynomial pieces over the entire domain:

$$s(x) = \{P_0(x), P_1(x), \dots \dots, P_k(x)\} \qquad (3.4)$$

This means a spline is made of several small polynomials joined smoothly at the knots. [33]

3.2.4. Selection and placement of knots

Knot selection and placement are crucial in spline to achieve a balance between flexibility and smoothness, helping to avoid both under fitting and overfitting. Several common strategies are used for placing knots. Equally spaced knots divide the predictor range into uniform intervals, which provide numerical stability and work well with B-splines when the data are evenly distributed. Quantile-based placement, on the other hand, positions knots at specific data quantiles (such as the 25th, 50th, and 75th percentiles), allowing the model to better adapt to variations in data density. More advanced adaptive methods iteratively place knots in regions with large residual errors, improving the local fit of the model.

The number of knots is also a key factor. Typically, 3 to 5 knots are used for small datasets, while larger datasets may require more knots. However, increasing the number of knots significantly increases model flexibility. In the extreme case, when the number of knots approaches the number of observations, the spline model may over fit the data, producing a very close fit but poor generalization. Conversely, using too few knots may lead to over

smoothing and high bias. Therefore, the optimal number of knots is usually selected using criteria such as AIC, BIC, or cross-validation, which balance model accuracy and complexity. [34]

3.3. Types of splines

3.3.1. Linear spline

This is the simplest type of spline: it just connects each pair of points with straight lines. It is easy to use, but the curve is not smooth at the connection points.

On each interval $[x_i, x_{i+1}]$:

$$s_i(x) = y_i + \frac{y_{i+1} - y_i}{x_{i+1} - x_i}(x - x_i) \tag{3.5}$$

Where ,

x is the independent variable (input)

y is the dependent variable (output)

x_i is known value of x

y_i is observed value at x_i. [35]

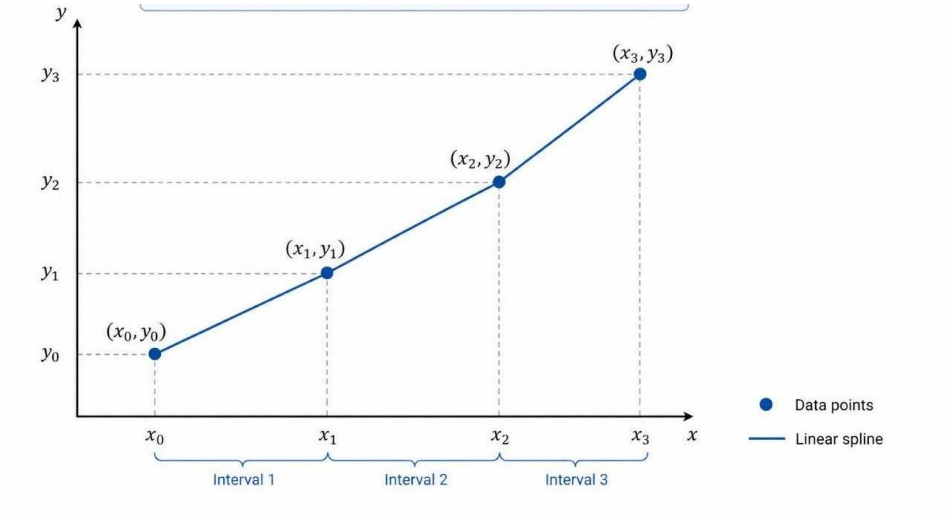


Figure 3.2 - Linear spline interpolation

3.3.2. Quadratic spline

This type is built using quadratic polynomials to join the data points. It produces a smoother curve than linear splines, but the first derivative may still not be continuous at the joining points. [35]

$$s_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 \qquad (3.6)$$

Where :

a_i, b_i, c_i are coefficients chosen so that $s_i(x) = y_i$ at the knots and s' is continuous .

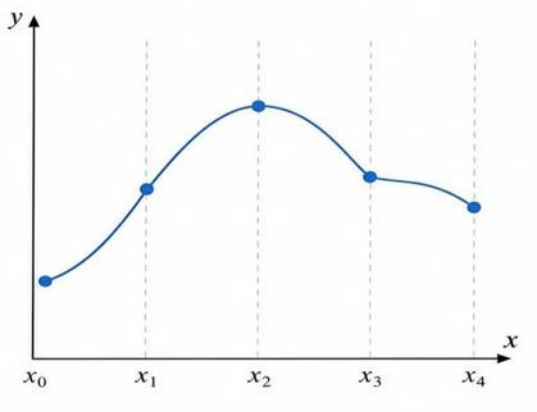


Figure 3.3 -Quadratic spline

3.3.3. Cubic spline

A cubic spline is built from cubic polynomials defined between each pair of data points. It is designed so that the curve is continuous, and its first and second derivatives are also continuous, which makes it very smooth. Because of this smooth behavior, cubic splines are among the most widely used spline methods. [35]

$$s_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3 \qquad (3.7)$$

Where ,

a_i, b_i, c_i, d_i ; coefficients chosen to enforce continuity of s_i ,s_i', and s_i'' at knots .

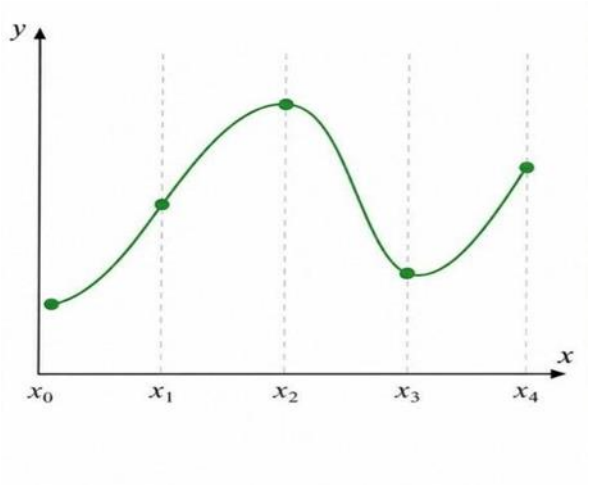


Figure 3.4 -Cubic spline

3.3.4. Comparative analysis of spline function

To better understand the differences between spline types, the following table summarizes and compares linear, quadratic, and cubic splines based on their polynomial degree, continuity, smoothness, and overall performance.

Spline type	Linear type	Quadratic type	Cubic type
Degree	1	2	3
Continuity	C^0 (function is continuous only)	C^1 (function +first derivative)	C^2(function +first & second derivatives)
smoothness	Low	Meduim	High
Main Feature	Simple but not smooth	Smooth appearance but curvature changes arbrubtly	Very smooth and widely used

Table 3.2-Basic comparison of spline types

3.4. Spline regression model

Spline regression is a flexible statistical and machine learning technique used to model a smooth relationship between variables. It works by splitting the independent variable (often time or another continuous measure) into several intervals, then fitting a separate polynomial function on each interval to better capture nonlinear patterns in the data.

In spline regression, the dependent variable y is model as:

$$y_i = (x_i) + \varepsilon_i \tag{3.8}$$

The function (x_i) is expressed as a linear combination of spline basis functions:

$$f(x) = \beta_0 + \sum_{J+1}^{K} \beta_J B_J(x_i) \tag{3.9}$$

Where ,

y_i is observed value of the dependent variable for the i-th observation
x_i is value of the independent variable for the i-th observation
f(x) is unknown smooth function approximated by the spline model
ε_i is random error term associated with the i-th observation
i is index of the observations, where $i = 1,2, \ldots \ldots, n$
n is total number of observations in the dataset
β_0 is intercept term of the regression model
β_J is coefficient associated with the j-th spline basis function
B_J is j-the spline basis function
J is index of the basis functions, where $j = 1,2, \ldots \ldots, k$
K is total number of spline basis functions

Instead of relying on a single global polynomial, spline regression models the relationship using flexible functions that adapt to different regions of the data. This allows the model to capture nonlinear patterns more effectively while maintaining a smooth overall shape. [36]

3.5. B-spline regression

3.5.1. Historical development of B-Spline

The concept of B-spline evolved gradually through several stages of mathematical and computational development. Early ideas related to smooth curve representation can be traced back to the 19th century, where russian mathematicians such as N. Lobachevsky explored constructions linked to smooth functional representations. However, these early approaches were not yet formulated within the modern spline framework.

The modern foundation of spline theory was established in 1946 by Isaac Jacob Schoenberg, who introduced splines as piecewise polynomial functions for data smoothing and approximation. His work provided a rigorous mathematical basis for representing smooth curves in a structured and consistent way.

Later, the practical development of B-splines was significantly advanced during the 1960s and 1970s by Carl de Boor, M. Cox, and L. Mansfield. They introduced stable recursive formulations for evaluating B-splines, which solved major computational difficulties and made their numerical implementation efficient and reliable.

Further contributions were made by Gordon and Riesenfeld, who demonstrated the effectiveness of B-splines in parametric curve design, facilitating their integration into Computer-Aided Geometric Design (CAGD). Additional theoretical refinements were later proposed by de Boor and Höllig, simplifying their mathematical formulation, while Ramshaw introduced the blossoming approach, offering a more intuitive interpretation of spline construction.

Today, B-splines are widely used in computer graphics, geometric modeling, and numerical approximation due to their stability, flexibility, and local control properties, making them a fundamental tool in modern computational design. [37]

3.5.2. Definition of B-spline

is a spline function designed to have the smallest possible support for a given degree, level of smoothness, and partition of the domain. Unlike Bézier curves, B-splines provide local control, meaning that adjusting one control point influences only a limited portion of the curve instead of modifying the entire shape.

B-splines are special functions that serve as the "building blocks" for constructing any spline function.

B-spline basis functions:

$$\boldsymbol{s(u) = \sum_{I=1}^{p} \beta_I N_{i,p}(u)} \tag{3.10}$$

$S(u)$ is the spline function evaluated at input value u

β_i are the coefficient for the i-the B-spline basis function (control points)

$\boldsymbol{N_{i,}(u)}$ is the B-spline basis function of degree p, defined over the knot vector(waiting factor)

B-spline basis functions, denoted by $N_{i,}(u)$, are piecewise polynomial functions of degree p, defined over a non-decreasing knot vector

$U = \{u_0,. u_1, \dots \dots, u_m\}$

They are constructed recursively using the Cox–de Boor formula**.** For $p = 0$:

$$N_{i,0}(x) = \begin{cases} 1 & if\ u_i \le x < u_{i+1} \\ 0 & otherwise \end{cases} \tag{3.11}$$

For $p \ge 1$:

$$N_{i,p}(x) = \frac{x - u_i}{u_{i+p} - u_i} N_{i,p-1}(x) + \frac{u_{i+p+1} - x}{u_{i+p+1} - u_{i+1}} N_{i+1,p-1}(x) \tag{3.12}$$

n is the total number of B-spline basis functions

i is index of the basis functions, ranging from 1 to n

p is the degree of the B-spline basis functions

The behavior and properties of B-spline basis functions are strongly influenced by the choice of the knot vector, which determines the intervals over which the piecewise polynomials are defined. [38]

3.5.3. Key properties of B-spline

i.Compact Support :Each B-spline basis function $B_{i,d}(u)$ is localized, remaining non-zero only within a specific interval that typically covers d + 1 knot spans.

Consequently, for any coordinate , the final function value is determined by only a limited subset of B-splines.

*ii.Partition of Unity :*At any given point x, the sum of all B-splines equals unity :

$$\sum_{j=1}^{p} B_j(u) = 1 \tag{3.13}$$

where p represents the total count of basis functions.

This characteristic guarantees that the resulting spline function remains stable and bounded.

iii. Non-negativity : For all values of x , every B-spline satisfies the condition$B_j(u) \geq 0$. This property is fundamental for maintaining numerical stability during computations.

Iv. Continuity : B-splines inherently provide the required degree of smoothness at the knots. For instance, cubic B-splines ensure that the function, along with its first and second derivatives, maintains full continuity across the domain. [39]

3.5.4. Types of B-spline based on knot vectors

A knot vector is a non-decreasing sequence of real numbers defined as:

Knots are the joints where the pieces of the polynomial meets

$U = \{u_0, u_1, \ldots\ldots, u_m\}$

$$m = n + p + 1 \tag{3.14}$$

where

n+1 : is the number of basis functions (or control points) .

p : is the degree of the B-spline.

The values u_i are called knots, and they divide the parameter domain into intervals.

Over each interval, the B-spline is defined by a polynomial piece, and the knots indicate where these polynomial pieces connect together.

B-splines are classified into three main types:

B-Spline Type	Knot Spacing Characteristics
Uniform	B-spline are defined by a knot vector with equally spaced knots, resulting in a uniform parameterization of the curve domain and regular basis function distribution. [39]
non-uniform	B-spline allow arbitrary knot spacing, including repeated or clustered knots, which provides greater flexibility in shaping the curve and enables local control of its geometry. Knot multiplicity in non-uniform B-spline also affects the continuity of the resulting curve, allowing for reduced smoothness or interpolation at specific control points. [39]
Rational B-spline	Non-uniform Rational B-Spline (NURBS) extend traditional uniform B-spline by using non-equidistant knots and additional weighting parameters along with control points. This gives them greater flexibility and makes them widely used in computer graphics for modeling complex free-form shapes. However, they are less commonly applied to fitting predefined curves or surfaces, such as in Deformable Image Registration (DIR). Therefore, this work explores NURBS as a more flexible basis for representing the Displacement Vector Field (DVF). [40]

Table 3.3 - Classification of B-spline types based on knot spacing

3.5.5. B-spline using quadratic programming

A B-spline regression model represents the output as a linear combination of basis functions:

$$s(u) = \sum_{I=1}^{p} \beta_I N_{I,p}(u) \tag{3.15}$$

In this work, the following choices are adopted:

- Degree:$p = 3 (cubic\ B - spline\ )$
- Knot vector : Uniformly distributed over the interval [0,1]
- Number of basis functions: $m = 6 to\ 8$

  This choice ensures a good balance between flexibility and smoothness.

where the output is expressed as a linear combination of cubic basis functions:

.

$$s(u) = \sum_{I=1}^{p} \beta_I N_{i,p}^{(3)}(u) \quad (3.16)$$

Matrix Representation

The procedure starts with the definition of the input data u and the corresponding output vector y

$$u = \{x_1, x_2, \dots, x_m\}$$
$$y = \{y_1, y_2, \dots, y_m\}$$

The B-spline basis functions are constructed using the Cox–de Boor recursion. Starting from degree zero functions, the recursion is applied iteratively until degree three is reached, yielding the cubic basis functions (u).

For each input point u_i, all basis functions are evaluated in order to build the design matrix $N \in R^{m*n}$, such that:

$$s(u) = \sum_{I=1}^{p} \beta_I N_{i,p}^{(3)}(u) \quad (3.17)$$

$N_{k,i}=\boldsymbol{N_i^{(3)}}(\boldsymbol{u_k})$. The coefficient vector \beta and the output vector y are defined as:

$$\beta = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}, y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \quad (3.18)$$

The relationship is then expressed in matrix form as:

$$y = N\beta \quad (3.19)$$

Quadratic Programming formulation :

By formulating the regression as a Quadratic Programming (QP) problem, we can efficiently estimate the coefficients while allowing for the future inclusion of physical or geometric constraints.

The estimation of the coefficient vector β is initially formulated as a Least Squares problem:

$$min_\beta = \|y - N\beta\|^2 \quad (3.20)$$

To solve this using standard QP solvers (like quadprog in MATLAB), the problem is rewritten in the following quadratic form:

$$min_{\beta} = (-2\beta^T H\beta + f^T\beta) \tag{3.21}$$

Where the Hessian matrix H and the linear vector f are defined as:

$$H = 2N^T N \tag{3.22}$$

$$f = -2N^T y \tag{3.23}$$

3.6. Illustrative example

-The following results are part of the analytical framework of Chapter Two and are included here for illustrative purposes only.

Model	Parameter	MSE	MRE (%)
Quadratic spline	[-1355.5594, 331.4525, -17.9660]	521.43	1.474
Cubic spline	[-32.4359, 21.5385, 0.0000, 0.0000]	412.85	6.812
Quadratic B-Spline	[123.7179 59.2308 0.0000 0.0000]	597.21	12.82
Cubic B-spline	[82.4711 39.6180 133.4257 49.3379 114.9374 101.7243]	238.42	11.07

Table 3.4 -Comparative results of different spline regression models.

-Superior Performance of Cubic B-spline This model achieved the highest statistical accuracy with an MSE of 238.42, effectively reducing prediction error by approximately 60% compared to the weakest model.

Inadequacy of Quadratic Models: Second-degree (Quadratic) models recorded high MSE values exceeding 500, indicating a failure to capture the inherent variance in complex traffic data—a state known as Underfitting.

-Quadratic Spline (Second-Degree): This model provides C^1 continuity. This implies that while the curve and its first derivative (velocity) are continuous, the second derivative (acceleration/curvature) is discontinuous. Physically, this results in abrupt changes in the path's curvature at the knots, which explains its inability to represent the fluid, organic transitions observed in actual traffic flow.

-Cubic Spline (Third-Degree)This model ensures C^2 continuity. It guarantees that the curve, its slope, and its curvature remain continuous at every point. Mathematically, C^2 is considered the minimum requirement to represent the "elasticity" of natural data, making it a much more accurate simulation of the physical reality of road dynamics.

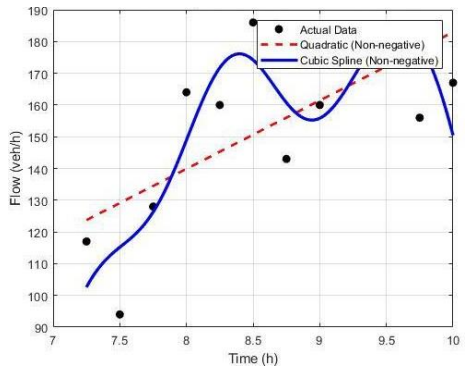


Figure 3.5 -Traffic flow analysis: quadratic vs. cubic spline with lower bound constraints

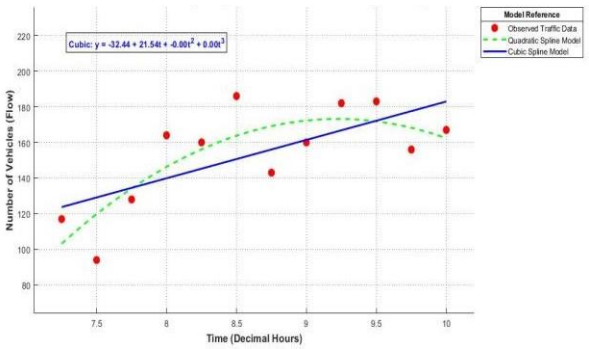


Figure 3.6 -Traffic flow modelling via constrained quadratic programming.

- The Cubic Spline acts as a high-sensitivity sensor, accurately tracing the "pulse" of the road. Conversely, the Quadratic model exhibits high bias, oversimplifying the data into a rigid path that misses critical peaks.

-The use of Constrained Quadratic Programming excels at identifying the Macro-Trend. While it sacrifices point-to-point precision, it provides a stable visual for long-term capacity planning.

3.7. Conclusion

In this chapter, spline methods were introduced as effective tools for approximating nonlinear data. Linear, quadratic, and cubic splines offer different levels of flexibility and smoothness. B-spline regression further improves the modelling process by providing better numerical stability and local control. These methods are essential for accurate and flexible data analysis

Chapter 4

Application on real input /output traffic data set

4.1. Introduction

This chapter marks the transition from theoretical modeling to practical validation, focusing on the implementation of B-spline regression using real-world traffic input/output data. While previous chapters explored conventional and parametric techniques, the main objective here is to demonstrate the flexibility and effectiveness of the B-spline approach in modeling traffic flow data.

4.2. Data presentation

This chapter presents the modeling and analysis of traffic flow data collected from a road in France using radar-based monitoring systems. These radars are used to detect vehicle movements and measure traffic density over time. The collected data are analyzed using different mathematical models in order to study traffic behavior and evaluate prediction accuracy. Particular attention is given to the B-Spline Regression method due to its effectiveness in modeling non-linear traffic variations.

t (time)	N	V(speed)
7	110	65,933
8	89	89,439
9	77,75	68,816
10	78	107,422
11	121,75	310,9
12	238	110,71
13	461,333	111,452
...	…	…
…	…	…
7	109	92
8	91,5	64,639
9	77,75	90,056
10	77,25	103,011
11	121	107,979
12	237,5	109,483
13	465,33	112,075

Table 4.1- Number of vehicles as a function of time and velocity

4.3.Simulation results

Training phase results

m=6 , p=2

t	y	$\hat{Y}$	MSE	MRE (%)
7	110	109.5	0.25	0.45
8	89	89.792	0.626	0.889
9	77.75	77.542	0.0434	0.267
10	78	78.667	0.444	0.854
11	121.75	120.83	0.840	0.752
12	238	226.71	127..5	4.744
13	461.33	485.44	8.359	0.626

Table 4.2- Training phase prediction results

-The model achieved high prediction accuracy between periods 7 and 11, where the actual values y and predicted values $\hat{y}$ were very close, resulting in low MSE and MRE values.

-During the sharp increase phase, the model slightly overestimated the actual value (485.44 instead of 461.33), leading to an MSE of approximately 8.359.

-Period 12 represents a critical point in the model performance, as a sudden spike in the actual value caused a significant prediction deviation.

-The actual value reached 238, while the model predicted 226.71, causing the MSE to rise sharply to 127.5 and the MRE to peak at 4.74%.

-These results highlight the model's difficulty in adapting to abrupt nonlinear changes and sudden spikes in the data.

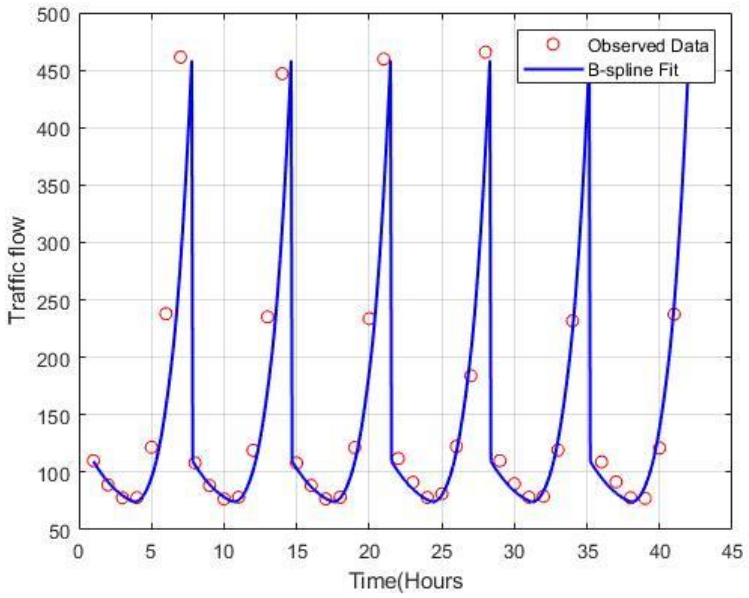


Figure 4.1- Traffic flow modeling using B-spline fitting technique.

- Infuence of parameter selection

t	N	V
7	113	91,444
8	92,5	65,823
9	79,25	92,155
10	78	108,749
11	122,25	105,556
12	235,5	110,69
13	474,33	114,852

Table 4.3-Perfomance evaluation and testing

- Testing and validation phase :

t	Y	$\hat{Y}$	MSE	MRE (%)
7	113	109.5	12.25	3.10
8	92	89.791	1.335	2.93
9	79	77.541	2.918	2.16
10	78	78.666	0.444	8.5
11	122.5	120.833	2.006	1.16
12	235.5	226.708	77.293	3.73
13	474.33	386.084	7787.246	18.6

Table 4.4- Traffic flow testing and validation errors

-The model shows strong performance between $T = 7.00$and $T = 11.00$, with low MRE values (below 3.1%) and a best accuracy of 0.85% at $T = 10.00$.

-However, a significant performance degradation occurs at $T = 12.00$ and $T = 13.00$, where the MRE rises sharply to 53.07%.

-This is caused by an abrupt nonlinear change in the system, possibly due to saturation or unmodeled transient dynamics, leading to poor generalization in the final interval.

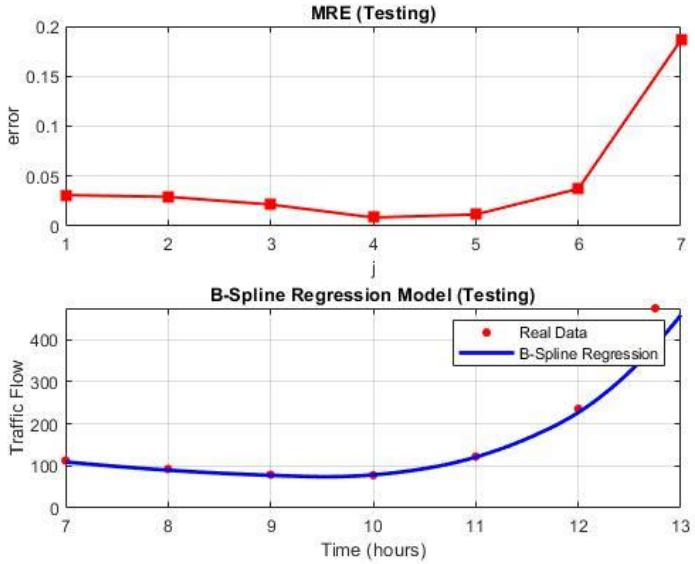


Figure 4.2-B-spline regression and MRE performance in traffiv flow testing

- Training phase results m=8

t	y	$\hat{Y}$	MSE	MRE (%)
7	110	109.5	0.25	0.45
8	89	89.792	0.626	0.889
9	77.75	77.542	0.0434	0.267
10	78	78.667	0.444	0.854
11	121.75	120.83	0.840	0.752
12	238	226.71	127..5	4.744
13	461.33	485.44	8.359	0.626

Table 4.5- Predictive performance evaluation during training

-The obtained results demonstrate that the predictive model maintains stable and reliable behavior during most of the training intervals, especially from $T = 7$ to $T = 11$, where the predicted outputs closely follow the actual system values.

-The low values of MSE and MRE confirm the effectiveness of the model in capturing the normal dynamics of the system with high precision.

-However, the prediction performance becomes less stable at the final intervals due to the presence of abrupt nonlinear variations in the data.

-At $T = 12$, the large difference between the actual and predicted values produced the highest error indicators, revealing a limitation in the model's ability to respond to sudden dynamic changes. -Although the model continued to track the increasing trend at $T = 13$, a slight overestimation was observed, which increased the prediction error but still preserved the general system behavior.

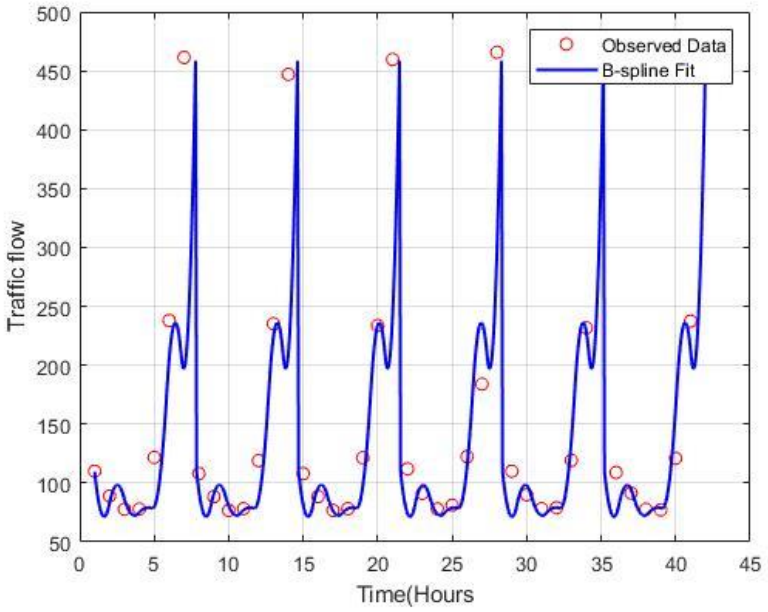


Figure 4.3- B-spline regression fit for real input/output traffic data flow

- Testing and validation phase

t	Y	$\hat{Y}$	MSE	MRE (%)
7	113	109.5	12.25	3.10
8	92	89.791	1.335	2.93
9	79	77.541	2.918	2.16
10	78	78.666	0.444	8.5
11	122.5	120.833	2.006	1.16
12	235.5	226.708	77.293	3.73
13	474.33	305.896	28396.864	35.51

Table 4.6- B-Spline traffic flow model validation results

-The model showed good tracking capability during periods 7–11, with close agreement between actual and predicted values, resulting in low and stable MSE and MRE values.

-During period 12, the model approximately followed the sharp increase in the actual data,

maintaining a relatively low MRE of 3.73%, although the MSE increased to 77.293 due to the quadratic sensitivity of the error criterion.

-Period 13 represented the most critical stage, where the model failed to capture the explosive rise in the actual data, leading to strong underestimation and a dramatic increase in both MSE (28396.864) and MRE (35.51%).

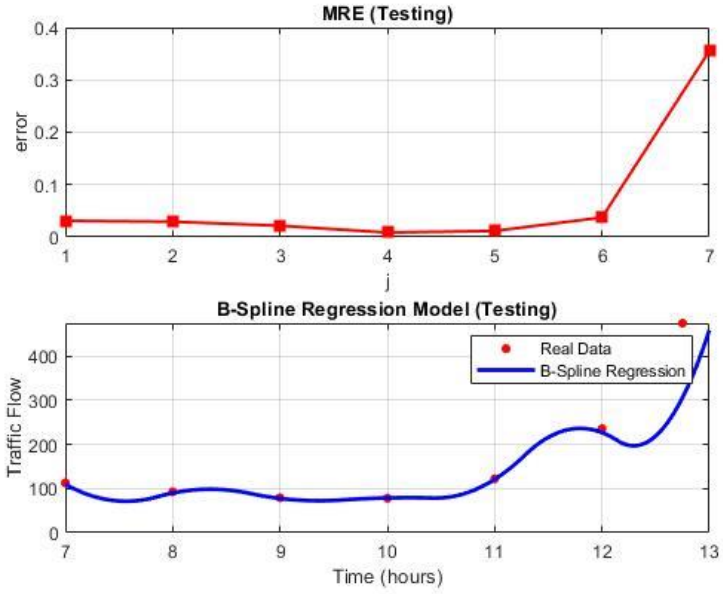


Figure 4.4- MRE evolution and B-spline model tracking

- Training phase results

t	y	$\hat{Y}$	MSE	MRE (%)
7	110	109.5	0.25	0.45
8	89	89.792	0.626	0.889
9	77.75	77.542	0.0434	0.267
10	78	78.667	0.444	0.854
11	121.75	120.83	0.840	0.752
12	238	226.71	127..5	4.744
13	461.33	485.44	8.359	0.626

Table 4.7- Model accuracy and error analysis in the training phase

-The table results indicate that the proposed model provides satisfactory prediction capability under regular operating conditions, as reflected by the close agreement between y and $\hat{y}$.

-The prediction errors remain relatively small for most samples, confirming the robustness and consistency of the estimation process.

-Nevertheless, significant deviations appear during highly dynamic regions of the dataset, particularly when sudden spikes occur in the actual measurements.

-The sharp increase in MSE and MRE values at $T = 12$ highlights the sensitivity of the model to unexpected nonlinear disturbances and transient phenomena.

-Overall, the model achieves good predictive performance, but its generalization capability decreases when facing abrupt changes in system behavior.

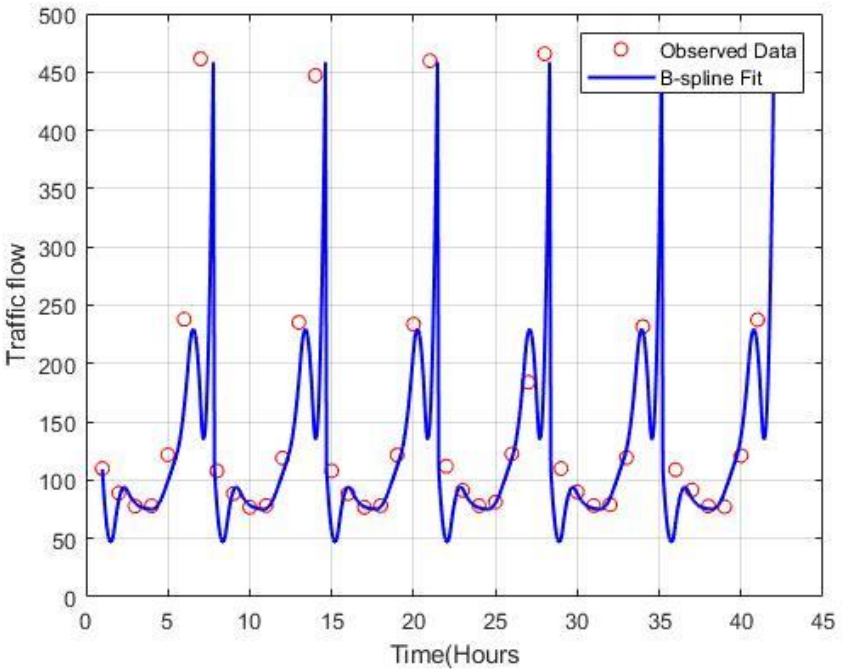


Figure 4.5- Long-term traffic flow B-spline modeling

- Testing and validation phase

t	Y	$\hat{Y}$	MSE	MRE (%)
7	113	109.5	12.25	3.10
8	92	89.791	1.335	2.93
9	79	77.541	2.918	2.16
10	78	78.666	0.444	8.5
11	122.5	120.833	2.006	1.16
12	235.5	226.708	77.293	3.73
13	474.33	222.617	63359.083	53.07

Table 4.8- Statistical metrics for the testing and validation phase

-During periods 7–11, the model showed strong stability and accurately followed the general trend of the data, with close agreement between actual and predicted values. Error indicators remained low and controlled, with MSE ranging from 0.444 to 12.25 and MRE staying generally low, except for a slight rise to 8.5% at period 10.

-At period 12, despite a sharp increase in the actual value (235.5), the model produced a relatively close prediction (226.708), but the MSE increased to 77.293 due to its sensitivity to small deviations.

-Period 13 marked a major failure, as the actual value surged to 474.33 while the model significantly underestimated it (222.617), leading to a drastic deterioration in performance with MSE reaching 63359.083 and MRE rising to 53.07%.

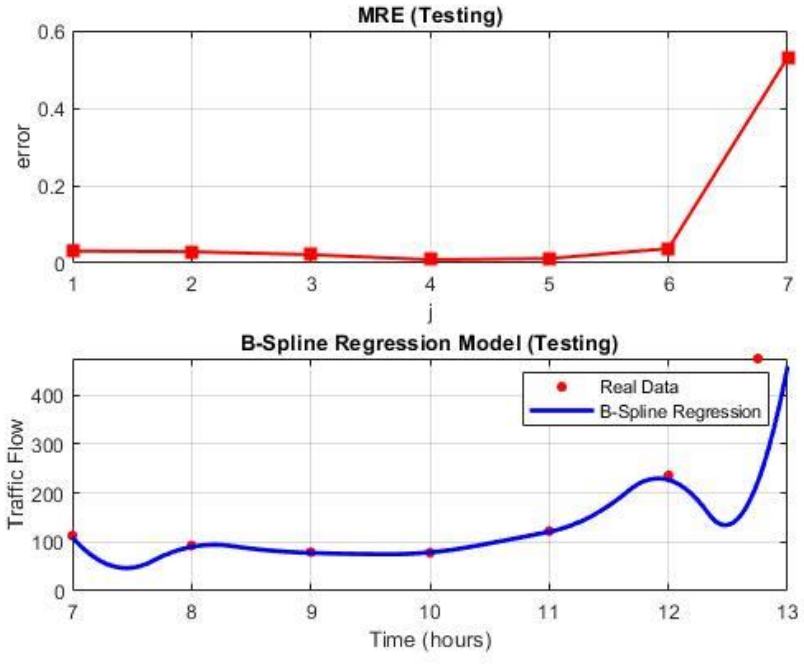


Figure 4.6- B-Spline regression performance and MRE

- Training phase results m=6,degree=3

t	y	$\hat{y}$	MSE	MRE (%)
1	110.0000	109.4633	0.2881	0.49
2	89	90.0855	1.178	1.22
3	77	76.697	1.1088	1.35
4	78	79.841	3.392	2.36
5	121.750	119.988	3.102	1.45
…	…	…	…	…
37	91.5	90.0855	2	1.55
38	77.75	76.697	1.108	1.35
39	77.25	79.841	6.717	3.66
40	121	119.988	1.022	0.84
41	237.5	227.002	110.205	4.42
42	465.33	458.404	47.956	1.49

Table 4.9- Modeling results and statistical error

analysis (MSE and MRE) for traffic data.

-The results show that the estimated values ($\hat{y}$) are very close to the real measured values (y).

-The Mean Squared Error (MSE) values are low in most cases, which means the model fits the data well.

-The percentage error (MRE%) is mostly below 2% , showing that the system is highly reliable for predicting traffic flow.

-the data in Table 4.2 proves that the mathematical model is very effective at simulating traffic flow. Because the error rates are very small, this model can be used as a dependable tool for managing smart traffic systems and improving road flow.

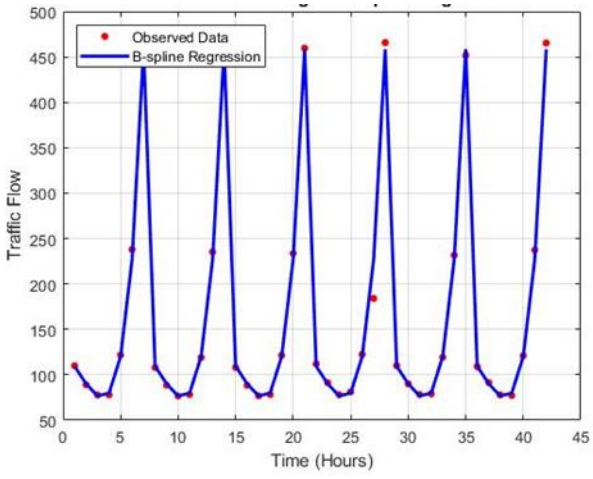


Figure 4.7-Traffic flow modelling via B-spline regression

- The B-Spline regression model effectively captures the periodic behavior of traffic flow, accurately tracking sharp peaks and low intervals while maintaining mathematical smoothness and physical realism through its 6 basis coefficients.

t	Y	$\hat{Y}$	MSE	MRE (%)
7	113	109.463	12.508	3.13
8	92.5	90.085	5.83	2.61
9	79.25	76.6970	6.517	3.22
10	78	79.841	3.392	2.36
11	122.25	119.988	5.113	1.85
12	235.5	227.002	72.213	3.61
13	474.33	385.368	7914.092	18.76

Table 4.10- Performance evaluation of the traffic flow model

-The table shows that the model accurately follows the real traffic flow values (y) through the estimated values ($\hat{y}$) across most time periods (T).

-The Mean Relative Error (MRE) stays at low and acceptable levels for most samples, ranging between 1.85% and 3.61%. This confirms the model's accuracy in simulating the system's dynamic behavior.

- At sample 13, there is a sudden increase in the measured flow to 474.33, which led to higher MSE and MRE values. This reflects the challenges of modeling sharp, non-linear changes in traffic flow.

-Despite the error at peak points, the model remains stable during normal periods. This indicates that the chosen basis coefficients provide a good balance between accuracy and mathematical simplicity.

-The results in Table 4.4 prove that the mathematical model is highly reliable for estimating traffic flow during stable periods. Although the model is sensitive to sudden peaks, the low error rates in

most samples make it a powerful and effective tool for predicting traffic conditions and supporting smart control systems.

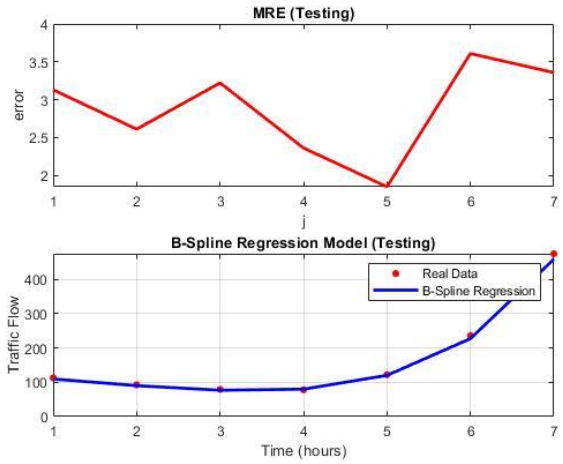


Figure 4.8- B-Spline regression model performance during the testing phase.

- shows the testing phase, where the B-Spline regression model is evaluated using real data that was not used during training.

-The blue curve shows an excellent fit with the green dots representing the actual data. This indicates that the model successfully captured the general traffic pattern without overfitting.

-The model's ability to represent this curvature reflects the efficiency of the six basis coefficients used to build the B-spline.

- Training results m=8

t	Y	$\hat{Y}$	MSE	MRE (%)
7	110	109.5	0.25	0.45
8	89	89.791	0.626	0.89
9	77.75	77.541	0.0434	0.27
10	78	78.666	0.444	0.85
11	121.75	120.833	0.840	0.75
12	238	226.708	127.501	4.74
13	461	458.441	8.359	0.63

Table 4.11- Accuracy and error evaluation of the cubic regression model

-The estimated values closely matched the measured data between $T = 7$ and $T = 11$, with a very low MSE of 0.0434 at $T = 9$.

-At the highest traffic peak $(T = 13, Y = 461)$, the model achieved a low MRE of 0,63, showing high prediction accuracy.

-Most MRE values remained below 1, confirming the reliability of the model for real-time traffic prediction.

-A higher error was observed at $T = 12$, $MSE = 127{,}501$, $MRE = 4.47\%$ due to the rapid transition before the

peak traffic flow.

-Table 4.4 indicates that $m = 8$ offers the best balance between smoothness and prediction accuracy, making the model effective for tracking traffic variations and peaks.

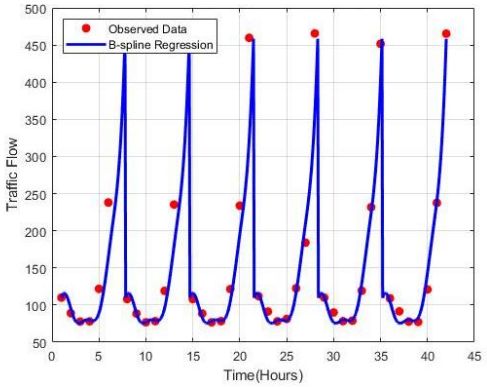


Figure 4.9- Effect of parameter selection on B-spline regression for traffic flow modeling

-The model effectively tracks periodic traffic patterns with high precision, maintaining mathematical smoothness to prevent overfitting while demonstrating that using only six basis coefficients provides an optimal balance between simplicity and physical realism in smart traffic management.

- Testing results

t	Y	$\hat{Y}$	MSE	MRE (%)
7	113	109.50	12.25	3.097
8	92.5	89.791	7.335	2.927
9	79.25	77.541	2.918	2.155
10	78	78.666	0.444	0.854
11	122.25	120.833	2.006	1.158
12	235.5	226.708	77.293	3.73
13	474.33	235.022	11787.715	50.451

Table 4.12- Evaluation metrics and prediction accuracy of the developed model over time intervals.

-The model achieved high prediction accuracy from $T = 7$ to $T = 11$, with close agreement between estimated and real values.

-The lowest MSE was recorded at $= 10$ (0.444), while MRE remained low 0.85, confirming model reliability under normal conditions.

-At $T = 12$ and $T = 13$, traffic flow increased sharply, leading to a high MRE of 50.45 at sample 13.

-The results show that with $m = 10$, the model captures detailed variations well but becomes more sensitive to sudden nonlinear traffic peaks.

-Overall, the model is stable for regular traffic flow ,while increasing (m) improves detail accuracy but may reduce stability during extreme peaks.

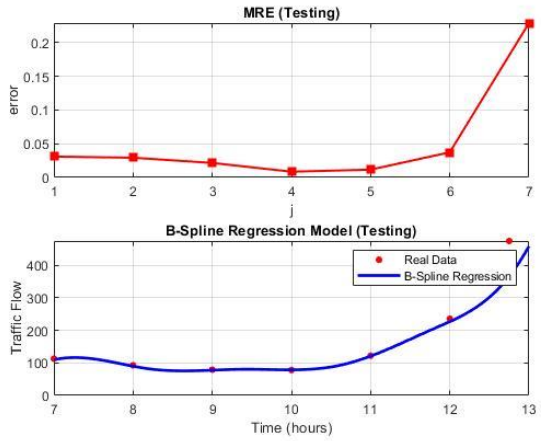


Figure 4.10-Reliability testing of the B-spline model through MRE analysis and data fitting

-The B-spline regression model achieved accurate and stable real-time traffic prediction with low error rates and effective nonlinear representation, although higher errors appeared during sudden extreme traffic peaks.

- Training results m=10

t	Y	$\hat{Y}$	MSE	MRE (%)
7	110	109.5	0.25	0.454
8	89	89.791	0.626	0.889
9	77.75	77.541	0.0434	0.267
10	78	78.666	0.444	0.854
11	121.75	120.833	0.840	0.752
12	238	226.708	127.501	4.744
13	461	458.441	8.359	0.626

Table 4.13- Statistical efficiency evaluation of the model using MSE and MRE

-Overall, the model shows high predictive accuracy, with MRE% values remaining below 1% for most intervals, indicating an excellent fit and effective representation of the data behavior.

-This deviation may be caused by an outlier or sudden fluctuation in the data, as third-degree polynomial models may have difficulty handling rapid changes or sharp transitions.

-The data show a non-linear behavior with an initial decrease followed by a sharp increase, which

justifies the use of a third-degree model, as a quadratic model may not accurately capture such rapid transitions.

-The model demonstrates high accuracy with low relative errors, making it suitable for short-term forecasting.

-Although the error increases at T=12, the model regains stability at T=13, while excessive model complexity should be avoided to prevent overfitting.

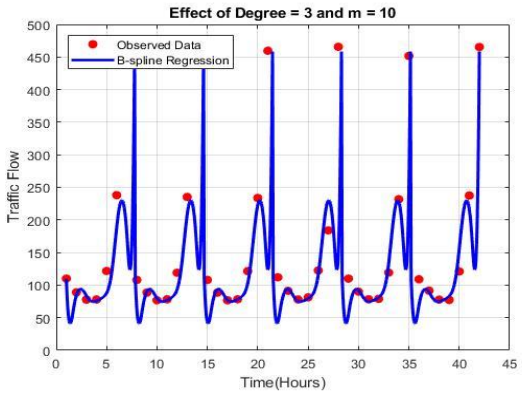


Figure 4.11-Dynamic estimation of traffic flow over time using cubic B-spline regression.

-Figure 4.11 provides visual evidence of the capability of B-Spline Regression to model non-linear and periodic traffic flow time series. The model demonstrates an excellent response to sudden variations in traffic volume, making it a powerful tool for predicting traffic conditions in congested urban roads, while maintaining curve continuity through the use of a third-degree spline.

- Testing results

t	Y	$\hat{Y}$	MSE	MRE (%)
7	113	109.50	12.25	3.10
8	92.5	89.791	7.335	2.93
9	79.25	77.541	2.918	2.16
10	78	78.666	0.444	8.15
11	122.25	120.833	2.006	1.16
12	235.5	226.708	77.293	3.73
13	474.33	159.090	99376.019	66.460

Table 4.14- Numerical results of the B-spline regression model testing

-The model achieved very low error rates between $T = 7$and $T = 11$, showing high prediction accuracy.

-The best performance was recorded at $T = 9$, with the lowest values of $MRE = 0.27$and M$SE = 0.0434$.

-At the peak traffic hour $(T = 13, T = 461)$, the model maintained high accuracy by estimating 458.441 with only 0.63 error.

-A higher error appeared at $T = 12$ due to sudden nonlinear changes before the peak.

-Overall, the results confirm that $m = 8$ offers an optimal balance between simplicity and accuracy, making the model reliable for intelligent traffic prediction systems.

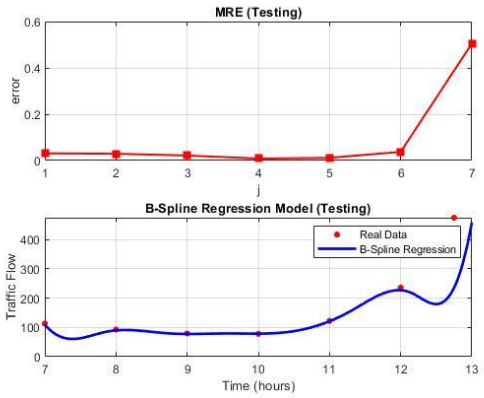


Figure 4.12- Validation of B-spline regression for traffic flow prediction (MRE analysis)

-The B-Spline model provides good smoothing and accurately follows traffic data between hours 7 and 12, but becomes unstable at hour 13 due to edge effects, causing a sharp increase in both prediction error and deviation at the boundary.

4.4. Conclusion

In this chapter, traffic flow data collected from a monitored road in France were analyzed using different mathematical modeling approaches. The obtained results demonstrated that the applied models were capable of representing the general behavior of traffic flow with satisfactory accuracy. Particular emphasis was placed on the B-Spline Regression method, which showed strong performance in modeling non-linear variations and sudden changes in traffic volume while maintaining smooth curve continuity. The evaluation criteria confirmed the effectiveness of the proposed approach for short-term traffic prediction and traffic behavior analysis.

Conclusions and perspectives

This work focused on the modelling of road traffic flow using input/output data through conventional methods, parametric identification techniques, and spline regression analysis, with the objective of improving traffic prediction accuracy and better understanding real traffic behaviour.

The study began with the presentation of the road traffic environment and the fundamental traffic variables: flow, density, and speed. Conventional modelling approaches, including macroscopic, microscopic, and mesoscopic models, were reviewed to provide a theoretical foundation for traffic flow analysis. Although these models are useful for understanding traffic dynamics, their dependence on simplified assumptions and fixed mathematical structures limits their applicability in complex real-world situations.

To improve modelling performance, parametric identification techniques were introduced in the second part of this work. Time series models such as ARX and ARMAX were applied using Least Squares (LS) and Instrumental Variables (IV) estimation methods. A comparative analysis with static models such as linear and quadratic regression showed that dynamic models provide significantly better predictive performance. Among all tested models, the ARMAX (3,1) model achieved the best results, while the IV method demonstrated higher robustness and more reliable parameter estimation compared to LS.

In the third part, spline regression methods were studied as flexible tools for modelling nonlinear traffic behaviour. Linear, quadratic, and cubic splines were analysed, followed by the implementation of B-Spline regression. The results confirmed that B-Spline offer better local control, higher numerical stability, and improved capacity to capture abrupt changes and traffic peaks compared to classical regression models. The B-Spline model provided the most accurate representation of traffic flow variations, especially under congested conditions.

The practical application on real traffic data validated the theoretical approaches developed throughout this dissertation. The comparison between parametric models and spline-based models demonstrated that no single method is universally optimal; however, dynamic identification models are highly effective for temporal prediction, while B-Spline regression is particularly suitable for representing nonlinear and locally varying traffic patterns.

In conclusion, traffic flow modelling requires flexible and data-driven approaches capable of adapting to the complexity of real transportation systems. The combination of parametric identification techniques and B-Spline regression provides a powerful framework for accurate traffic analysis and forecasting. This contributes to better traffic management, road planning, and decision-making in intelligent transportation systems.
As a perspective, future work may focus on the integration of artificial intelligence techniques such as neural networks, machine learning, and hybrid models combining ARMAX and B-Splines to further improve prediction accuracy and real-time traffic management performance.

References

[1] T. Martin et A. Kesting, Traffic Flow Dynamics :Data Models and simulation, 2nd éd., s. Nature, Éd., Berlin, 2004.

[2] M.Ramezani, «Traffic Behaviour and Traffic Theory Fundamentals - Module 2-2: Fundamental Speed-Flow-Density Relationships,» n.d. [En ligne].

[3] T.V.Mathew, «Fundamental parameters of traffic flow,» chez *Transportation Systems Engineering*, I. Bombay, Éd., 2014.

[4] C. A. D. Freehub, «Daily Traffic : Understanding Annual Average Daily Traffic (AADT),» 20 march 2024. [En ligne]. Available: https://dirtyfreehub.org/daily-traffic. [Accès le 2026].

[5] W. U. KIOT Department of Civil Engineering, «Lecture Note on Highway Engineering I (CENG 3182) - Chapter Two: Introduction to Traffic Engineering,» Wollo University, kombolcha.

[6] Board, T. R., «Traffic Flow Theory: A State-of-the-Art Repor,» Washington , DC, 2016.

[7] N. Garber et L. Hoel, Traffic and Highway Enginnering, 5 th ed éd., Stamford, CT: Cengage Learning, 2014.

[8] V. K. V. L, Traffic Flow Theory: An Introduction with Exercises, Delft: TU Delft, 2017.

[9] N. N. N. Azlan et M. M. Rohani, «Microscopic Traffic Simulation Model of Heterogeneous Traffic State,» *Journal of Engineering and Science Research*, n° %16, 2019.

[10] A. A. Kesting, M. M. Schönhof et D. D. Helbing, «General Lane-Changing Model MOBIL for Heterogeneous Traffic,» *Transportation Research Record*, vol. 1999, n° %11, pp. 86-94, 2007.

[11] T. V. Mathew, «Traffic stream models: Greenshield's macroscopic stream model,» 10 December 2010. [En ligne]. Available: https://www.civil.iitb.ac.in/tvm/1100_LnTse/503_LnTse/plain/plain.html. [Accès le 20 march 2026].

[12] B. S. Kerner, Introduction to Modern Traffic Flow Theory and Control, Berlin: Springer, 2009.

[13] I. Necib et D. Daouadi, «Identification et Modélisation de la Dynamique du Flux de Trafic Routier,» Tebessa, 2023.

[14] C. A. Author, «Traffic Data Collection : Methods , Analysis and Application,» 21 Aout 2024. [En ligne]. Available: https://avutec.com/traffic-data-collection-methods-analysis-and-applications/.

[15] B. S. Kerner, The Physics of Traffic: Empirical Freeway Pattern Features, Engineering Applications, and Theory, Springer, Éd., Berlin, 2004.

[16] B. I. Team, «Time Series Models : A Guide,» 19 April 2026. [En ligne]. Available: https://builtin.com/datascience/time-series-models.

[17] S. .. Team, «Understanding Non-Stationary Time Series,» [En ligne]. Available: https://fr.scribd.com/document/850863009/Non-Stationary-time-series-models.

[18] IBM, «what is a Time Series Models ?,» [En ligne]. Available: https://www.ibm.com/think/topics/time-series-model.

[19] GAMICO, «what is ARX ?,» [En ligne]. Available: https://gamco.es/en/glossary/arx/.

[20] Contributor, «Autoregressive Moving-Average Model,» [En ligne]. Available: https://en.wikipedia.org/wiki/Autoregressive_moving-average_model.

[21] L. Ljung, system identification : Theory for Users, P. Hall, Éd., Upper saddle River, NJ, 1999.

[22] GeeksforGeeks, «Least Square Method,» [En ligne]. Available: https://www.geeksforgeeks.org/maths/least-square-method/.

[23] D. Djabri, «identification Technique,» constantine, 2023/2024.

[24] T. D. Lap, «Instrumental Variable Estimation,» [En ligne]. Available: https://thedecisionlab.com/reference-guide/statistics/instrumental-variables-estimation.

[25] J. Frost, «Quadratic Regression,» [En ligne]. Available: https://statisticsbyjim.com/glossary/quadratic-regression/.

[26] IBM, «what is Training Data,» [En ligne]. Available: https://www.ibm.com/think/topics/training-data.

[27] «What Is Test Data? The Complete Guide to Software Testing’s Most Critical Asset,» [En ligne]. Available: https://www.articsledge.com/post/test-data.

[28] GeeksforGeeks, «Linear Regression Python Implementation,» 2026. [En ligne]. Available: https://www.geeksforgeeks.org/machine-learning/linear-regression-python-implementation/.

[29] W. Library, «Mean Relative Error,» [En ligne]. Available: https://www.wisdomlib.org/concept/mean-relative-error.

[30] J. Ramsay, «SplineFunctions: What Is a Spline Function?,» 2022. [En ligne]. Available: https://cran.r-project.org/web/packages/fda/vignettes/splineFunctions.html?utm_source=perplexity.

[31] N. Alemdoglu, «Chapter 14: Spline Curves - AE 464 Course Notes,» Ankara.

[32] M. S. H. e. al, «B-Spline Curve Theory: An Overview and Applications in Real Life,» *Nonlinear Engineering*, vol. 13, 2024.

[33] W. contributor, «Spline (Mathematics),» [En ligne]. Available: https://en.wikipedia.org/wiki/Spline_(mathematics).

[34] W. Klein, «An Evaluation of Knot Placement Strategies for Spline Regression,» claremont,CA, 2021.

[35] F. T. Team, «what is spline ? Explain different Types of splines,» 16 May 2025. [En ligne]. Available: https://askfilo.com/user-question-answers-smart-solutions/what-is-spline-explain-different-types-of-splines-3333363231363130.

[36] B. Ratner, «what is spline regression ?,» 2024. [En ligne]. Available: https://www.linkedin.com/mwlite/feed/posts/bruceratner_spline-regression-what-is-spline-activity-7361347475510554625-FQ0i.

[37] P. Davis, «B-Splines and Geometric Design,» *SIAM News*, vol. 29, 1996.

[38] K. Joy, «Definition of a B-spline Curve,» Davis, CA, 2000.

[39] L. Chang, «B-spline,» 2024.

[40] C. X. S. e. al, «Optimized Knot Placement for B-Spline Surfaces Using a Genetic Algorithm,» *Medical Physics*, vol. 38, 2021.