Transport and Communications Science Journal

# ADVANCED PREPROCESSING OF GNSS DATA FOR TRAFFIC ANALYSIS USING THE HAMPEL FILTER, HYBRID INTERPOLATION MODELS, AND THE KALMAN FILTER

**Do Van Manh***

University of Transport and Communications, No 3 Cau Giay Street, Hanoi, Vietnam

ARTICLE INFO

TYPE: Research Article
Received: 04/01/2026
Revised: 25/02/2026
Accepted: 05/03/2026
Published online: 15/04/2026
https://doi.org/10.47869/tcsj.77.3.8

* *Corresponding author*
Email: manhdv@utc.edu.vn; Tel: +84 972573673

**Abstract.** GNSS-mounted data collected from vehicles has become a critical source for urban traffic research and management due to its ability to continuously capture position, speed, and trajectory in real time, thereby supporting the estimation of traffic density, average speed, travel time, congestion detection, and mobility behavior analysis. However, in urban environments, GNSS accuracy is often degraded by multipath effects, satellite signal obstruction caused by high-rise buildings, and device limitations, leading to noisy speed measurements, outliers, and missing data, which can significantly bias results if raw data are used directly. This study proposes a three-stage preprocessing framework: (i) applying a Hampel filter to detect and replace outliers based on the median; (ii) reconstructing missing data using conditional linear interpolation for short gaps, combined with neighboring value propagation methods (LOCF and NOCB) for boundary segments; and (iii) employing a Kalman filter to smooth the speed time series and eliminate near-zero values caused by GNSS noise. The results demonstrate that the proposed approach effectively restores a continuous speed profile, reduces measurement noise, and enhances data reliability for subsequent urban traffic analysis.

**Keywords:** ITS, GIS, GNSS, Hampel Filter, Kalman Filter, Urban Transportation.

**Tạp chí Khoa học Giao thông vận tải**

# TIỀN XỬ LÝ NÂNG CAO DỮ LIỆU GNSS CHO PHÂN TÍCH GIAO THÔNG BẰNG HAMPEL FILTER, MÔ HÌNH NỘI SUY KẾT HỢP VÀ KALMAN FILTER

**Đỗ Văn Mạnh***

Trường Đại học Giao thông vận tải, Số 3 Cầu Giấy, Hà Nội, Việt Nam

THÔNG TIN BÀI BÁO

CHUYÊN MỤC: Công trình khoa học
Ngày nhận bài: 04/01/2026
Ngày nhận bài sửa: 25/02/2026
Ngày chấp nhận đăng: 05/03/2026
Ngày xuất bản Online: 15/04/2026
https://doi.org/10.47869/tcsj.77.3.8

* *Tác giả liên hệ*
Email: manhdv@utc.edu.vn; Tel: +84 972573673

**Tóm tắt.** Dữ liệu GNSS gắn trên phương tiện đang trở thành nguồn thông tin quan trọng trong nghiên cứu và quản lý giao thông đô thị nhờ khả năng ghi nhận liên tục vị trí, vận tốc và quỹ đạo theo thời gian thực, qua đó hỗ trợ ước lượng mật độ phương tiện, tốc độ trung bình, thời gian hành trình, phát hiện ùn tắc và phân tích hành vi di chuyển. Tuy nhiên, trong môi trường đô thị, độ chính xác của GNSS bị suy giảm do đa đường dẫn, che khuất vệ tinh và hạn chế thiết bị, dẫn đến nhiễu vận tốc, giá trị ngoại lai và thiếu dữ liệu, gây sai lệch nếu sử dụng trực tiếp dữ liệu thô. Nghiên cứu đề xuất quy trình tiền xử lý ba bước gồm: (i) sử dụng bộ lọc Hampel để phát hiện và thay thế các giá trị ngoại lai dựa trên trung vị; (ii) tái tạo dữ liệu thiếu bằng nội suy tuyến tính có điều kiện cho các đoạn ngắn, kết hợp phương pháp lan truyền giá trị lân cận (LOCF, NOCB) cho các đoạn đầu và cuối; (iii) áp dụng bộ lọc Kalman để làm mượt chuỗi vận tốc và loại bỏ các giá trị gần 0 do nhiễu. Kết quả cho thấy phương pháp giúp tái tạo chuỗi dữ liệu liên tục, giảm nhiễu và nâng cao độ tin cậy cho phân tích giao thông đô thị.

**Từ khóa:** ITS, GIS, GNSS, Hampel filter, Kalman filter, Giao thông đô thị.

## 1. ĐẶT VẤN ĐỀ

Hệ thống định vị vệ tinh toàn cầu (Global Navigation Satellite System- GNSS) đã trở thành một trong những nguồn dữ liệu quan trọng trong nghiên cứu và quản lý giao thông [1, 2]. Thông qua việc ghi nhận liên tục vị trí, vận tốc và thời gian của phương tiện, dữ liệu GNSS cho phép tái tạo quỹ đạo di chuyển, ước lượng mật độ phương tiện, tính toán tốc độ trung bình, phát hiện ùn tắc và đánh giá hiệu quả các giải pháp phân luồng trong đô thị. Với chi phí thu thập thấp và phạm vi quan sát rộng, dữ liệu GNSS ngày càng được sử dụng rộng rãi trong các hệ thống giao thông thông minh [3, 4].

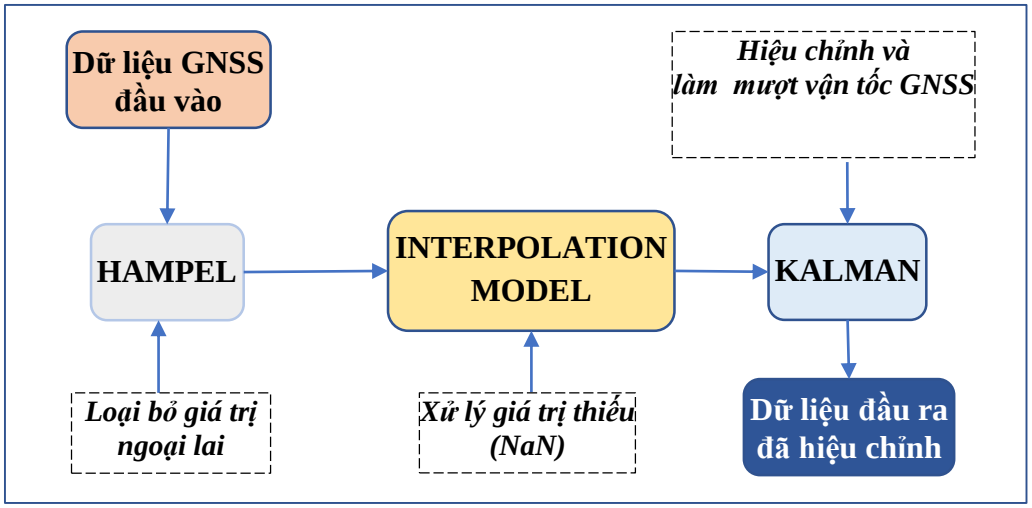
Tuy nhiên, chất lượng dữ liệu GNSS trong môi trường đô thị thường bị suy giảm đáng kể. Nguyên nhân chủ yếu bao gồm hiện tượng che khuất vệ tinh bởi nhà cao tầng, cây xanh và các công trình hạ tầng, dẫn đến mất tín hiệu hoặc giảm độ chính xác [5]. Hiệu ứng đa đường dẫn (multipath effect) khiến tín hiệu bị phản xạ trước khi đến anten thu, tạo ra giá trị vận tốc sai lệch lớn. Số lượng vệ tinh quan sát được (visible satellites) thay đổi liên tục theo thời gian, đặc biệt trong khu vực đô thị, làm tăng sai số tính toán vận tốc. Ngoài ra, các thiết bị giám sát hành trình (GPS onboard units) với tần suất ghi thấp hoặc độ phân giải hạn chế cũng tạo ra nhiều bản ghi không ổn định. Kết quả là chuỗi vận tốc GNSS thường xuất hiện giá trị nhiễu đột biến, vận tốc âm hoặc quá lớn, cùng các đoạn dữ liệu bị thiếu NaN (Not a Number) kéo dài [6].

Nếu sử dụng trực tiếp dữ liệu GNSS thô, các sai lệch này có thể làm sai lệch phân bố vận tốc, gây sai số trong ước lượng mật độ phương tiện, tốc độ trung bình hoặc phân loại trạng thái di chuyển. Các đoạn NaN làm đứt quãng chuỗi vận tốc, trong khi các giá trị gần bằng 0 xuất hiện do nhiễu tín hiệu có thể dẫn đến đánh giá sai về tình trạng tắc đường. Do đó, việc xây dựng một quy trình tiền xử lý dữ liệu GNSS toàn diện, ổn định và có khả năng khôi phục chuỗi vận tốc liên tục là yêu cầu bắt buộc trước khi tiến hành các bước phân tích giao thông [7].

Nhằm giải quyết các vấn đề trên, nghiên cứu này trình bày một quy trình tiền xử lý nâng cao gồm ba bước: (i) sử dụng Hampel filter để phát hiện và loại bỏ giá trị ngoại lai vận tốc lớn bất thường; (ii) khôi phục dữ liệu vận tốc bị thiếu thông qua mô hình nội suy kết hợp bao gồm Conditional Linear Interpolation (CLI), Last Observation Carried Forward (LOCF) và Next Observation Carried Backward (NOCB); (iii) áp dụng Kalman filter để hiệu chỉnh trạng thái chuyển động và loại bỏ các vận tốc nhỏ do nhiễu tín hiệu. Quy trình đề xuất giúp tái tạo chuỗi vận tốc GNSS đầy đủ, giảm nhiễu đo và nâng cao độ tin cậy của dữ liệu phục vụ phân tích giao thông đô thị.

## 2. PHƯƠNG PHÁP NGHIÊN CỨU

Phương pháp nghiên cứu được xây dựng nhằm cải thiện chất lượng dữ liệu vận tốc GNSS trước khi đưa vào các mô hình phân tích giao thông. Quy trình gồm ba bước chính (Hình 1): (i) phát hiện và loại bỏ giá trị ngoại lai bằng Hampel filter; (ii) khôi phục toàn bộ dữ liệu vận tốc bị thiếu bằng mô hình nội suy kết hợp; và (iii) hiệu chỉnh trạng thái chuyển động bằng Kalman filter. Chuỗi xử lý này giúp dữ liệu GNSS trở nên ổn định, liên tục và phù hợp cho các bước phân tích chuyên sâu ở giai đoạn sau.




Hình 1. Quy trình tiền xử lý nâng cao dữ liệu GNSS từ thiết bị giám sát hành trình gắn trên phương tiện giao thông có đăng ký kinh doanh vận tải.

### 2.1. Lọc ngoại lai bằng Hampel filter

Hampel filter là phương pháp phát hiện ngoại lai dựa trên trung vị, nhằm loại bỏ các giá trị đột biến trong chuỗi thời gian mà không làm thay đổi xu hướng dữ liệu [8, 9]. Với mỗi điểm dữ liệu xi, một cửa sổ trượt *Wi* có kích thước *2k+1* được xét:

$$W_i = \{x_{i-k}, \ldots\ldots, x_i, \ldots\ldots, x_{i+k}\} \quad (1)$$

Trung vị trong cửa sổ được tính theo:

$$m_i = median(W_i) \quad (2)$$

Độ lệch tuyệt đối trung vị (MAD) được xác định bằng:

$$MAD_i = median(|x_j - m_i|), j \in W_i \quad (3)$$

MAD được chuẩn hóa để ước lượng độ lệch chuẩn:

$$\sigma_i = 1.4826\ MAD_i \quad (4)$$

Một điểm được coi là ngoại lai nếu thỏa điều kiện:

$$|x_i - m_i| > n.\sigma_i \quad (5)$$

Trong đó: hệ số *n* là ngưỡng phát hiện ngoại lai của bộ lọc Hampel, dùng để xác định mức độ sai lệch cho phép của một quan sát so với trung vị. Giá trị *n* càng lớn thì điều kiện phát hiện ngoại lai càng nghiêm ngặt. Trong nghiên cứu này, *n* được lựa chọn theo khuyến nghị phổ biến trong tài liệu tham khảo để cân bằng giữa việc loại bỏ nhiễu và bảo toàn xu thế dữ liệu.

Khi đó, giá trị bị thay thế bằng trung vị:

$$x_i^* = m_i \quad (6)$$

Ngược lại, nếu không vượt ngưỡng, dữ liệu giữ nguyên giá trị ban đầu.

Hampel filter được áp dụng để loại bỏ các giá trị vận tốc bất thường sinh ra do nhiễu GNSS, đa đường, hoặc lỗi thiết bị. Bộ lọc hoạt động bằng cách so sánh từng điểm vận tốc với trung vị của nhóm điểm lân cận và đánh dấu các giá trị có sai lệch quá lớn. Các điểm ngoại lai được thay thế bằng trung vị, giúp chuỗi vận tốc trở nên ổn định hơn, giảm nhiễu đột biến và giữ

nguyên xu thế chuyển động của phương tiện. Điều này tạo điều kiện thuận lợi cho các bước xử lý tiếp theo như nội suy dữ liệu thiếu và lọc Kalman.

### 2.2. Khôi phục dữ liệu thiếu bằng mô hình nội suy kết hợp CLI với LOCF và NOCB

Sau khi loại bỏ các giá trị bất thường, dữ liệu GNSS vẫn còn xuất hiện nhiều đoạn NaN do mất tín hiệu. Để tái tạo chuỗi vận tốc đầy đủ theo thời gian, mô hình nội suy kết hợp được áp dụng gồm ba thành phần.

$$V_i = \begin{cases} V_i & \text{Nếu } v_i \neq NaN \\ V_{i-1} & \text{Nếu } V_i = NaN \text{ và tồn tại } V_{i-1} \neq NaN \quad (LOCF) \\ V_{i+1} & \text{Nếu } V_i = NaN \text{ và tồn tại } V_{i+1} \neq NaN \quad (NOCB) \\ V_{t_i} = V_{k-1} + \frac{V_{m+1} + V_{k-1}}{t_{m+1} + t_{k-1}} (t_i - t_{k-1}) ; & \forall i \in [k,m] \text{ Nếu Nếu } V_i = NaN \text{ và tồn tại } V_{k-1};\ V_{k+1} \text{ tồn tại. } (CLI) \end{cases} \quad (7)$$

CLI được sử dụng cho các đoạn dữ liệu bị thiếu nằm giữa hai giá trị vận tốc hợp lệ liên tiếp. Phương pháp này nội suy tuyến tính theo thời gian nhằm đảm bảo sự liên tục và xu hướng vận tốc không bị thay đổi đột ngột. CLI đặc biệt hiệu quả đối với các khoảng mất tín hiệu ngắn.

Đối với các đoạn NaN nằm ở đầu chuỗi, cuối chuỗi hoặc khoảng mất tín hiệu dài không thỏa điều kiện của CLI, hai kỹ thuật LOCF và NOCB được áp dụng. LOCF thay thế NaN bằng giá trị hợp lệ gần nhất phía trước, trong khi NOCB sử dụng giá trị hợp lệ gần nhất phía sau. Hai kỹ thuật này giúp tránh ngoại suy tuyến tính và giữ ổn định chuỗi vận tốc [10].

Xử lý các trường hợp đặc biệt như trường hợp một phương tiện chỉ có duy nhất một giá trị vận tốc hợp lệ, toàn bộ chuỗi được gán bằng giá trị này. Nếu phương tiện không có bất kỳ giá trị hợp lệ nào, vận tốc được thay bằng giá trị trung bình của toàn bộ tập dữ liệu. Điều này đảm bảo không còn giá trị NaN trong bất kỳ chuỗi nào.

### 2.3. Hiệu chỉnh vận tốc bằng Kalman filter

Kalman filter được sử dụng để làm mượt chuỗi vận tốc GNSS và loại bỏ các giá trị nhiễu [11, 12], đặc biệt là các vận tốc gần bằng 0 km/h phát sinh do tín hiệu không ổn định trong môi trường đô thị. Bộ lọc mô tả vận tốc dưới dạng mô hình trạng thái tuyến tính:

$$x_k = x_{\{k-1\}} + w_k,\ z_k = x_k + v_k \quad (8)$$

Trong đó xk là vận tốc thực, zk là vận tốc đo từ GNSS, còn wk và vk lần lượt là nhiễu quá trình và nhiễu đo.

Quá trình ước lượng gồm hai bước:

**(i) Dự đoán**

$$\hat{x}_{\{k|k-1\}} = \hat{x}_{\{k-1|k-1\}};\ P_{\{k|k-1\}} = P_{\{k-1|k-1\}} + Q \quad (9)$$

**(ii) Cập nhật**

$$K_k = \frac{P_{\{k|k-1\}}}{P_{\{k|k-1\}} + R} \quad (10)$$

$$\hat{x}_{\{k|k\}} = \hat{x}_{\{k|k-1\}} + K_k \left( z_k - \hat{x}_{\{k|k-1\}} \right) \quad (11)$$

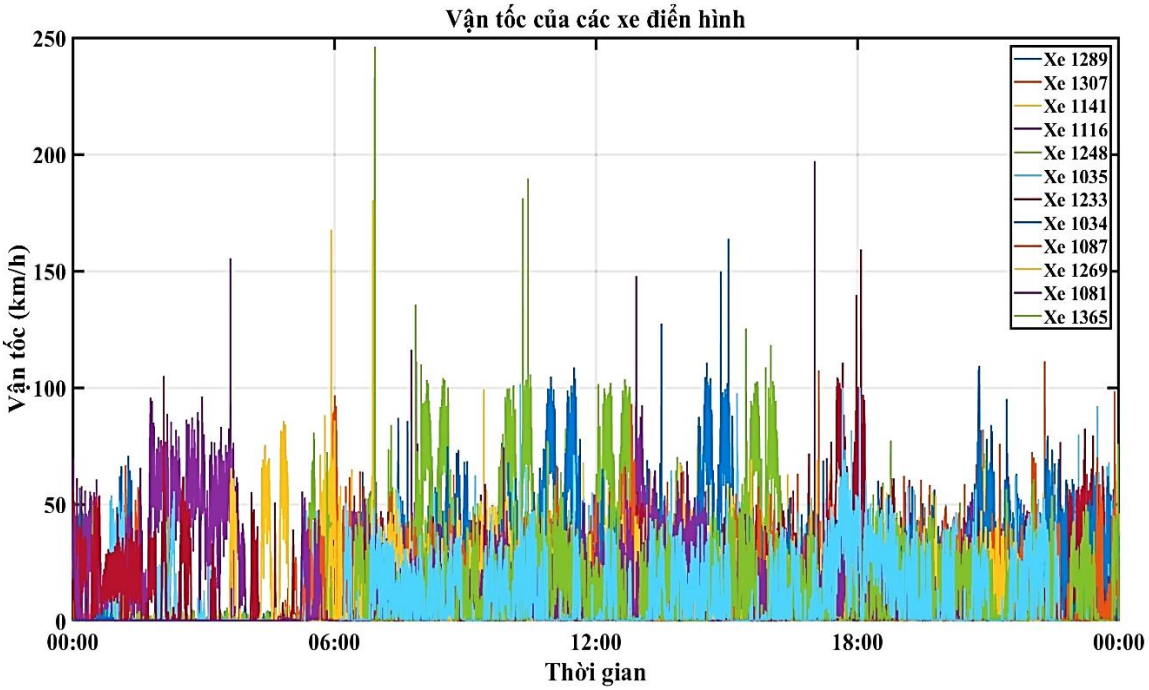
Kalman filter giúp giảm nhiễu đo đột biến, hiệu chỉnh các vận tốc 0 km/h không thực, và khôi phục chuỗi vận tốc liên tục, ổn định. Đây là bước cuối trong quy trình tiền xử lý nhằm

nâng cao chất lượng dữ liệu GNSS trước khi phân tích bài toán về phân cụm xác định điểm tắc nghẽn giao thông trong khu vực đô thị.

## 3. THỰC NGHIỆM

### 3.1. Đặc điểm dữ liệu

Bộ dữ liệu được thu thập từ hệ thống định vị vệ tinh toàn cầu GNSS (Global Navigation Satellite System) gắn trên các phương tiện giao thông hoạt động trong khu vực đô thị. Dữ liệu gồm hơn 1.048.000 bản ghi được ghi nhận liên tục theo thời gian từ một đơn vị cung cấp thiết bị giám sát hành trình trên thành phố Hà Nội, với mỗi dòng dữ liệu chứa các trường thông tin: thời gian (timestamp), mã định danh phương tiện, kinh độ, vĩ độ, và vận tốc tính toán (km/h). Tổng cộng có 153 phương tiện mẫu tham gia trong bộ dữ liệu này.




Hình 2. Tệp mẫu vận tốc theo thời gian của một số xe đăng ký kinh doanh vận tải điển hình.

Qua phân tích sơ bộ, dữ liệu phản ánh đa dạng các trạng thái chuyển động của phương tiện, với vận tốc trung bình khoảng 15,94 km/h. Tuy nhiên, cũng ghi nhận một số bất thường như giá trị vận tốc cực đại lên tới ~280 km/h, 20 điểm dữ liệu vượt quá 200 km/h, và 156 điểm bị thiếu giá trị vận tốc. Đáng chú ý, không có dòng nào có vận tốc bằng 0, song tồn tại khoảng 26.690 điểm có vận tốc rất nhỏ (< 0,1 km/h), có thể là trạng thái xe dừng nhưng bị nhiễu vị trí GNSS (Hình 2).

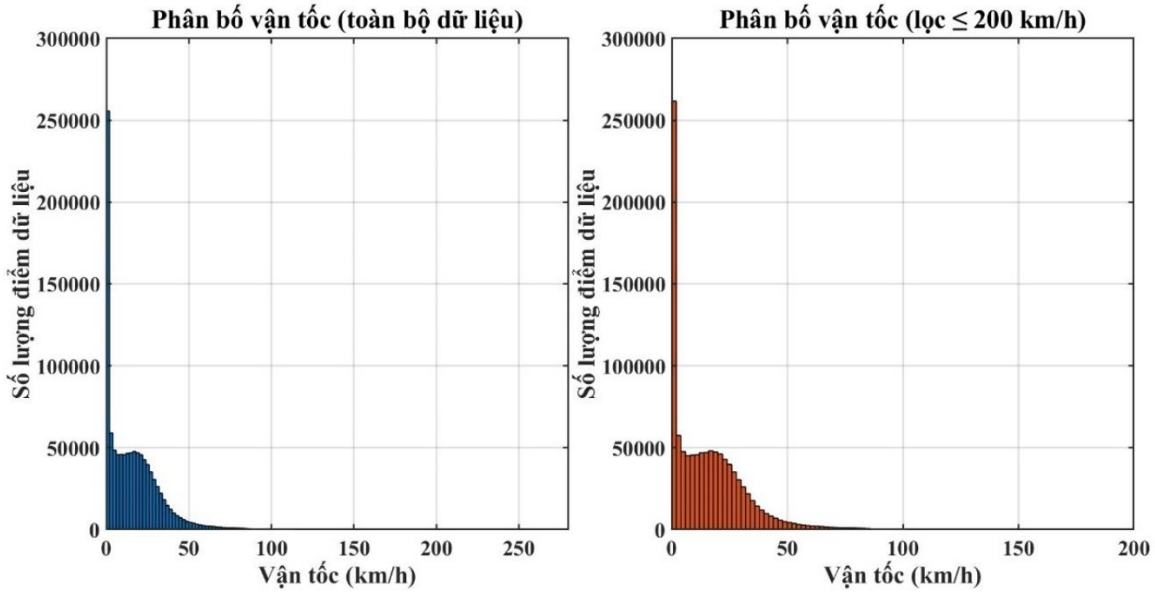
Bộ dữ liệu này là cơ sở quan trọng cho các bước tiền xử lý, làm sạch dữ liệu và phân tích theo khung thời gian nhằm nhận diện các điểm nóng giao thông đô thị, chi tiết ở Bảng 1.

Bảng 1. Thống kê mô tả chất lượng dữ liệu vận tốc GNSS của xe đăng ký kinh doanh vận tải-lỗi tín hiệu và cách xử lý nâng cao.

| STT | Chỉ số | Giá trị | Ghi chú cảnh báo | Phương pháp lọc nhiễu GPS / xử lý lỗi tín hiệu |
|---|---|---|---|---|
| 1 | Tổng số dòng dữ liệu | 1.048.575 | | Không áp dụng (thống kê tổng thể). |
| 2 | Số lượng xe khảo sát | 153 xe | | Không áp dụng (chỉ thống kê số lượng phương tiện). |
| 3 | Vận tốc trung bình | ~15,94 km/h | Trong khoảng hợp lý | Không áp dụng (thống kê tổng thể). |
| 4 | Số dòng có vận tốc > 200 km/h<br>(Vận tốc lớn nhất 280.35 km/h) | 20 | Có thể là lỗi GPS hoặc lỗi tính toán tự động của thiết bị Giám sát hành trình | Phát hiện và loại bỏ điểm bất thường bằng Hampel Filter |
| 5 | Số dòng bị thiếu vận tốc (NaN) | 156 | Thiếu dữ liệu cần loại bỏ hoặc xử lý nội suy | Xử lý dữ liệu bị thiếu (NaN) phương pháp Conditional Linear Interpolation (CLI) để khôi phục chuỗi dữ liệu. |
| 6 | Vận tốc rất nhỏ (<0,1 km/h) | 26.690 điểm<br>(không có giá trị = 0) | Có thể là trạng thái xe dừng nhưng bị nhiễu vị trí GNSS | Đối với các giá trị vận tốc rất nhỏ (< 0,1 km/h), Kalman Filter được áp dụng để loại bỏ nhiễu và xác định chính xác liệu phương tiện đang dừng hay tín hiệu GNSS đang bị sai lệch. |

## 3.2. Kết quả tiền xử lý dữ liệu

### 3.2.1. Phát hiện và loại bỏ điểm vận tốc lớn bất thường bằng Hampel Filter

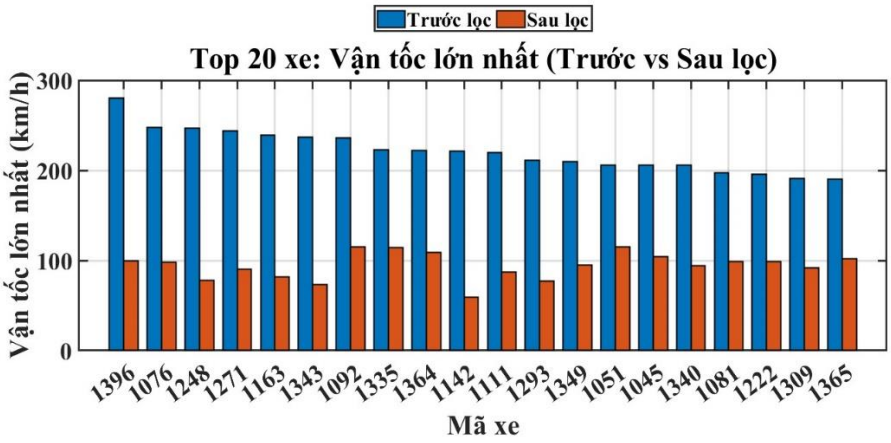



(a) (b)

Hình 3. Biểu đồ phân bố vận tốc của tệp dữ liệu: (a) Toàn bộ dữ liệu; (b) lọc vận tốc cao bất thường bằng Hampel Filter (V > 200 km/h).

Trong tiêu chuẩn nhận dạng ngoại lai của Hampel, nghiên cứu này sử dụng ngưỡng $n = 3$ (tương ứng khoảng $3\sigma$ với $\sigma$ ước lượng từ MAD chuẩn hóa), đây là lựa chọn phổ biến trong

phát hiện ngoại lai bền vững (robust) nhằm loại bỏ các đột biến vận tốc do nhiễu GNSS (Hình 3) mà vẫn giữ xu thế động học tổng thể của chuỗi vận tốc. Kết quả mẫu 20 xe trước và sau khi lọc được thể hiện trong Hình 4.




Hình 4. Nhóm 20 xe có vận tốc lớn nhất (trước và sau khi lọc).

Bảng 2. Kết quả phân tích tốc độ phương tiện trước và sau lọc ngoại lai bằng phương pháp Hampel.

| Data | Tổng điểm bất thường | Vận tốc lớn nhất trước lọc (Km/h) | Vận tốc lớn nhất sau lọc (Km/h) | Vận tốc trung bình sau lọc (Km/h) | Tỷ lệ % vượt ngưỡng trước lọc (%) | Tỷ lệ % vượt ngưỡng Sau lọc (%) | Số xe |
|---|---|---|---|---|---|---|---|
| 1048575 | 48662 | 280,350 | 119,796 | 15,744 | 0,421 | 0,355 | 153 |

Trong đó:

$$\text{Tỷ lệ \% vượt ngưỡng trước lọc (\%)} = \frac{\text{Số data có tốc độ >80 km/h}}{\text{Tổng số data}} \times 100\% = \frac{4410}{1048575} \times 100\% = 0{,}421\ \% \quad (12)$$

$$\text{Tỷ lệ \% vượt ngưỡng sau lọc (\%)} = \frac{\text{Số data (đã lọc) có tốc độ >80 Km/h}}{\text{Tổng số data}} \times 100\% = \frac{3720}{1048575} \times 100\% = 0{,}355\% \quad (13)$$

Bảng 2 ở trên trình bày các chỉ số tổng quan của dữ liệu tốc độ phương tiện sau khi lọc ngoại lai bằng thuật toán Hampel. Tổng cộng có 1.048.575 bản ghi từ 153 xe, trong đó 48.662 điểm bị phát hiện là bất thường và đã được xử lý. Sau lọc, vận tốc cực đại giảm từ 280,35 km/h xuống 119,79 km/h và vận tốc trung bình còn 15,74 km/h, phản ánh dữ liệu thực tế hơn. Đồng thời, tỷ lệ vượt ngưỡng tốc độ cũng giảm từ 0,421% xuống còn 0,355%, cho thấy dữ liệu sau xử lý chính xác và tin cậy hơn để phục vụ phân tích.

Trong thuật toán Hampel được áp dụng trong nghiên cứu này, việc thay thế các giá trị vận tốc bất thường không được thực hiện một cách ngẫu nhiên mà dựa trên một đại lượng thống kê có độ ổn định cao, trung vị (median). Cụ thể, khi một điểm dữ liệu được xác định là ngoại lai, thuật toán không loại bỏ hoàn toàn điểm đó mà thay thế bằng giá trị trung vị được tính toán từ một "cửa sổ trượt" gồm các quan sát lân cận có kích thước xác định. Trong nghiên cứu này, cửa sổ trượt được thiết lập với bán kính k= 7, nghĩa là tổng cộng 15 điểm dữ liệu (bao gồm 7 điểm trước, điểm hiện tại và 7 điểm sau) được xem xét khi đánh giá mỗi quan sát. Giá trị trung

vị được xác định bằng cách sắp xếp 15 giá trị vận tốc trong cửa sổ theo thứ tự tăng dần và chọn giá trị nằm ở vị trí trung tâm của dãy đã sắp xếp. Phương pháp thay thế này giúp duy trì đặc trưng động học của chuỗi vận tốc, giảm thiểu tác động của các biến động bất thường và đảm bảo tính liên tục của dữ liệu mà không làm sai lệch kết quả trong các phân tích thống kê tiếp theo.

### 3.2.2. Xử lý dữ liệu bị thiếu (NaN) bằng kết hợp CLI với LOCF và NOCB

*Input:*
*Bảng T với các cột:*
*MaXe : mã xe*
*ThoiGian : thời gian*
*Speed_Filtered_km_h: vận tốc sau Hampel (có thể có NaN)*
*Output:*
*Speed_CLI_km_h : vận tốc sau khi điền đầy*
*IsCLIInterpolated : cờ đánh dấu các giá trị được điền (TRUE/FALSE)*
*Thuật toán:*
*1. Sắp xếp T theo (MaXe, ThoiGian).*
*2. Gán:*
*speed0 = T.Speed_Filtered_km_h*
*veh = T.MaXe*
*time = T.ThoiGian*
*speed_CLI = speed0 // bản sao để cập nhật*
*IsCLIInterpolated = FALSE cho mọi dòng*
*3. Tính vận tốc trung bình toàn bộ:*
*globalMean = giá trị trung bình của speed0 (bỏ qua NaN)*
*4. Với mỗi xe vID trong tập các giá trị khác nhau của MaXe:*
*Lấy chỉ số idx các dòng có MaXe = vID*
*v = speed0(idx) // chuỗi vận tốc của xe vID*
*t = time(idx) // chuỗi thời gian tương ứng*
*N = chiều dài v*

*Xác định:*
*validMask = (v không phải NaN)*
*nValid = số phần tử TRUE trong validMask*
*Nếu nValid = 0:*
*// xe không có vận tốc hợp lệ*
*gán v(i) = globalMean cho mọi i = 1..N*
*đánh dấu IsCLIInterpolated(idx(i)) = TRUE*
*Ngược lại nếu nValid = 1:*
*// chỉ có 1 điểm vận tốc*
*v_single = giá trị duy nhất của v tại validMask = TRUE*
*với mọi i = 1..N nếu v(i) là NaN:*
*gán v(i) = v_single*
*IsCLIInterpolated(idx(i)) = TRUE*
*Ngược lại (nValid ≥ 2):*

Để xử lý các bản ghi thiếu vận tốc (NaN), nghiên cứu sử dụng phương pháp Conditional Linear Interpolation (CLI). Trước tiên, chuỗi vận tốc của từng phương tiện được sắp xếp theo thời gian và quét để xác định các đoạn NaN liên tiếp. Với mỗi đoạn thiếu dữ liệu, tổng thời gian gián đoạn được tính theo biểu thức:

$$\Delta T = \sum_{i=k}^{m-1}(t_{i+1} - t_i) \quad (14)$$

Trong đó: k và m lần lượt là chỉ số bắt đầu và kết thúc của đoạn bị thiếu vận tốc. Nếu tổng thời gian gián đoạn thỏa mãn điều kiện:

$$\Delta T \leq T_{max} \quad (15)$$

Với $T_{max}$= 5 phút (ngưỡng được lựa chọn dựa trên đặc thù mật độ và tần suất thu thập dữ liệu GNSS của nghiên cứu này), và đồng thời tồn tại giá trị vận tốc hợp lệ ngay trước ($v_{k-1}$) và ngay sau đoạn thiếu ($v_{m+1}$), vận tốc tại các điểm trong đoạn NaN được ước lượng bằng nội suy tuyến tính theo thời gian (CLI) theo công thức (7).

Trong trường hợp ΔT > 5 phút hoặc không có điểm hợp lệ bao quanh, các giá trị NaN được giữ nguyên nhằm tránh tạo ra các ước lượng không phản ánh đúng động học của phương tiện. Các điểm này được loại khỏi những phân tích phụ thuộc vào vận tốc ở các bước tiếp theo. Phương pháp CLI cho phép khôi

```
        4.1 Chuyển t sang dạng số:
           t_num   = chuyển ThoiGian sang số
           t_valid = t_num(validMask)
           v_valid = v(validMask)
        4.2 Nội suy tuyến tính cho NaN nằm giữa:
           tạo v_mid = v
           với mọi i có v(i) = NaN và t_num(i) nằm
trong [min(t_valid), max(t_valid)]:
              v_mid(i) = nội suy tuyến tính từ (t_valid,
v_valid)
           (các NaN ở ngoài khoảng trên giữ nguyên)
        4.3 Forward-fill (LOCF) cho NaN phía sau:
           v_ffill = v_mid
           với i từ 2 đến N:
              nếu v_ffill(i) = NaN và v_ffill(i-1) ≠ NaN:
                 v_ffill(i) = v_ffill(i-1)
        4.4 Backward-fill (NOCB) cho NaN phía
trước:
           v_bfill = v_ffill
           với i từ N-1 xuống 1:
              nếu v_bfill(i) = NaN và v_bfill(i+1) ≠
NaN:
                 v_bfill(i) = v_bfill(i+1)
        4.5 Cập nhật kết quả:
           với mọi i = 1..N:
              nếu v(i) ban đầu là NaN và v_bfill(i)
không phải NaN:
                 speed_CLI(idx(i))      = v_bfill(i)
                 IsCLIInterpolated(idx(i)) = TRUE
5. Gán:
     T.Speed_CLI_km_h    = speed_CLI
     T.IsCLIInterpolated = IsCLIInterpolated
6. Sử dụng Speed_CLI_km_h làm chuỗi vận tốc đã
làm sạch
  cho các bước phân tích tiếp theo (phân cụm, mật
độ, v.v.).
```

phục chính xác các khoảng thiếu dữ liệu ngắn, đồng thời hạn chế sai lệch khi gặp các gián đoạn dài.

Trong quá trình thu thập dữ liệu GNSS, các bản ghi vận tốc có thể xuất hiện giá trị thiếu (NaN) do mất tín hiệu hoặc che khuất vệ tinh. Để tái tạo chuỗi vận tốc liên tục phục vụ phân tích, nghiên cứu sử dụng hai phương pháp bổ sung nhau: nội suy tuyến tính có điều kiện (CLI) cho các đoạn thiếu ngắn có điểm bao quanh, và phương pháp mở rộng CLI kết hợp LOCF–NOCB để xử lý các trường hợp mất dữ liệu dài hoặc nằm ở đầu/cuối chuỗi dữ liệu.

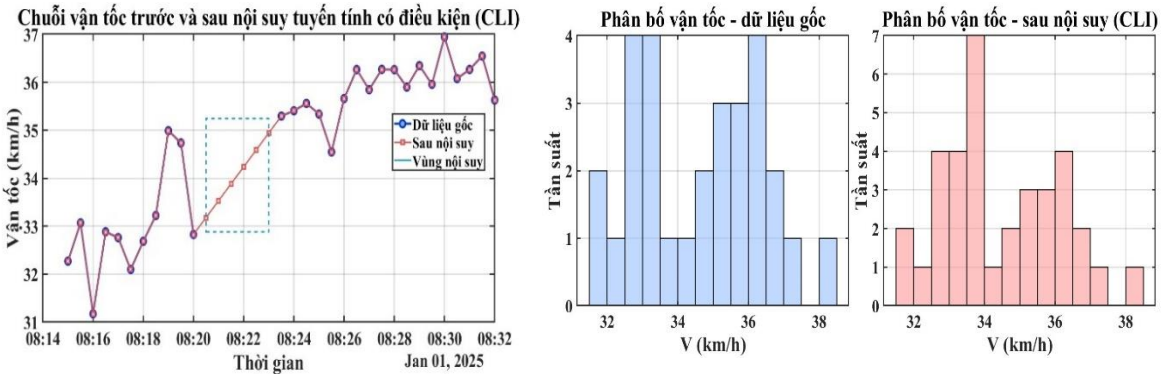
Mô hình gộp cho phép xử lý toàn bộ giá trị vận tốc bị thiếu bằng cách ưu tiên nội suy tuyến tính (CLI) khi đoạn NaN nằm giữa hai quan sát hợp lệ, và áp dụng LOCF hoặc NOCB cho các điểm thiếu ở đầu hoặc cuối chuỗi. Cách tiếp cận kết hợp này giúp tái tạo đầy đủ chuỗi vận tốc GNSS mà không tạo ra sai lệch lớn, đảm bảo dữ liệu đầu vào ổn định cho các phân tích tiếp theo. Chi tiết các bước xử lý nằm ở Hình 5.

Hình 5. Pseudo code xử lý dữ liệu bị thiếu (NaN).

Kết quả thực nghiệm cho thấy đặc trưng phân bố vận tốc gần như không thay đổi sau khi áp dụng CLI, chứng tỏ phương pháp này duy trì tính ổn định thống kê và đảm bảo chất lượng dữ liệu đầu vào cho các bước phân tích điểm nóng giao thông.

Kết quả cho thấy thuật toán nội suy tuyến tính có điều kiện (CLI) hoạt động ổn định và phù hợp với đặc trưng của chuỗi vận tốc GNSS. Trên Hình 6, ở biểu đồ (6a), đoạn dữ liệu bị thiếu được khôi phục mượt mà, bám sát xu hướng vận tốc lân cận mà không làm thay đổi hình dạng chung của chuỗi. Khung nét đứt giúp xác định rõ phạm vi nội suy.

Hai biểu đồ phân bố vận tốc bên phải cho thấy sự khác biệt trước và sau nội suy là rất nhỏ. Phân bố sau CLI gần như trùng với dữ liệu gốc, đồng thời không xuất hiện giá trị bất thường. Điều này khẳng định CLI chỉ bổ sung thông tin tại các đoạn thiếu ngắn mà không làm sai lệch đặc tính thống kê.

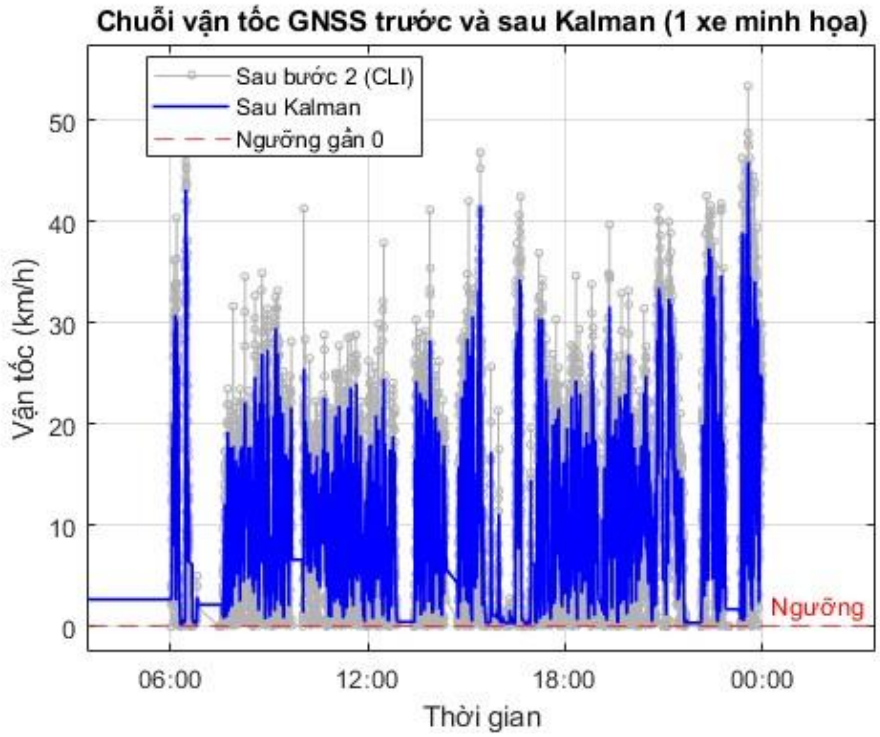



(a) (b)

Hình 6. So sánh dữ liệu vận tốc trước và sau nội suy tuyến tính có điều kiện (CLI): (a) Chuỗi vận tốc theo thời gian; (b) Phân bố vận tốc tương ứng.

Nhìn chung, phương pháp CLI giúp tái tạo chuỗi vận tốc liên tục, bảo toàn cấu trúc động học của dữ liệu và đảm bảo độ tin cậy cho các phân tích tiếp theo như phân cụm giao thông, ước lượng mật độ hay mô hình hóa dòng xe.

### 3.2.3. Khử nhiễu và làm mượt vận tốc GNSS bằng Kalman Filter




Hình 7. So sánh dữ liệu vận tốc trước và sau khi lọc Kalman.

Trong Hình 7 bên dưới cho thấy Kalman Filter đã làm mượt chuỗi vận tốc GNSS một cách hiệu quả, đặc biệt tại các đoạn mà dữ liệu xuất hiện nhiều dao động nhỏ hoặc các giá trị gần bằng 0 km/h do nhiễu tín hiệu. Đường vận tốc sau áp dụng Kalman trở nên ổn định hơn, giảm đáng kể các biến động bất thường và bám sát xu hướng chuyển động thực của phương tiện. Việc loại bỏ hoàn toàn các vận tốc 0 giả (không phải trạng thái dừng thật) cho thấy Kalman Filter hoạt động hiệu quả trong việc hiệu chỉnh dữ liệu vận tốc và nâng cao độ tin cậy của chuỗi

GNSS.

Kết quả chạy chương trình cho thấy việc áp dụng Kalman Filter đối với chuỗi vận tốc GNSS sau khi đã xử lý ngoại lai và nội suy đem lại hiệu quả rất cao trong việc làm sạch dữ liệu. Toàn bộ 26.690 điểm có vận tốc gần bằng 0 km/h do nhiễu tín hiệu GNSS đã được loại bỏ, phản ánh khả năng hiệu chỉnh mạnh của bộ lọc đối với các giá trị đo không thực.

Trong Bảng 3, độ lệch chuẩn giảm từ 15,264 xuống 13,916 km/h cho thấy chuỗi vận tốc sau lọc ổn định hơn và ít dao động bất thường hơn. Giá trị trung bình vận tốc hầu như không thay đổi (tăng 0,36%), chứng tỏ Kalman Filter làm mượt dữ liệu nhưng không làm sai lệch vận tốc thật của phương tiện bảng bên dưới.

Bảng 3. Thống kê chất lượng vận tốc trước và sau khi sử dụng phép lọc Kalman.

| STT | Chỉ tiêu | Trước Kalman | Sau Kalman | Thay đổi |
|---|---|---|---|---|
| 1 | Tổng số xe | 153 | 153 | Không |
| 2 | Số điểm vận tốc gần 0 (≤ 0,1 km/h) | 26.690 | **0** | Giảm hoàn toàn |
| 3 | Tỷ lệ điểm gần 0 | 2,545% | **0%** | Giảm 100% |
| 4 | Giá trị trung bình (km/h) | 15,744 | 15,801 | + 0,36% |
| 5 | Độ lệch chuẩn (km/h) | 15,264 | **13,916** | - 8,8% |

Phân bố vận tốc sau lọc cũng trở nên trơn tru và hợp lý hơn, phù hợp cho các phân tích giao thông tiếp theo như nhận dạng trạng thái chạy dừng, phân cụm, hoặc mô hình hóa lưu lượng. Nhìn chung, kết quả chứng minh rằng việc bổ sung bước Kalman sau Hampel và nội suy là cần thiết và mang lại cải thiện rõ rệt về chất lượng dữ liệu GNSS.

## 4. KẾT LUẬN VÀ KIẾN NGHỊ

Nghiên cứu đã đề xuất và triển khai một quy trình tiền xử lý nâng cao cho dữ liệu vận tốc GNSS gồm ba bước: (i) phát hiện và loại bỏ giá trị ngoại lai bằng Hampel Filter, (ii) tái tạo dữ liệu thiếu thông qua mô hình nội suy kết hợp, và (iii) làm mượt và hiệu chỉnh vận tốc bằng Kalman Filter. Kết quả thực nghiệm cho thấy chuỗi vận tốc sau tiền xử lý trở nên liên tục, ổn định và ít nhiễu hơn, đồng thời loại bỏ hoàn toàn các giá trị vận tốc gần bằng 0 km/h không phản ánh trạng thái di chuyển thực của phương tiện. Độ lệch chuẩn giảm đáng kể và phân bố vận tốc trở nên hợp lý hơn, chứng minh hiệu quả của quy trình trong việc nâng cao chất lượng dữ liệu GNSS.

Quan trọng hơn, dữ liệu GNSS sau khi tiền xử lý đã đảm bảo đủ độ tin cậy để sử dụng trong các tác vụ phân tích nâng cao, đặc biệt là phân cụm trạng thái giao thông, phát hiện bất thường, ước lượng mật độ phương tiện và nhận dạng tắc nghẽn đô thị. Việc cải thiện chất lượng dữ liệu đầu vào giúp các mô hình phân tích vận hành ổn định hơn, giảm sai số và phản ánh chính xác hơn hiện trạng giao thông, góp phần nâng cao hiệu quả giám sát, quản lý và điều hành giao thông đô thị.

Mặc dù quy trình tiền xử lý đề xuất cho thấy hiệu quả trong việc cải thiện chất lượng chuỗi vận tốc GNSS, nghiên cứu vẫn còn một số hạn chế. Việc đánh giá hiệu quả chủ yếu dựa trên các chỉ số thống kê và phân tích xu hướng, trong khi chưa có dữ liệu tham chiếu độc lập để

kiểm chứng độ chính xác tuyệt đối. Bên cạnh đó, mô hình Kalman sử dụng giả định động học tuyến tính, có thể chưa phản ánh đầy đủ các hành vi chuyển động phức tạp trong môi trường giao thông đô thị.

Trong các nghiên cứu tiếp theo, phương pháp có thể được mở rộng bằng cách tích hợp thêm thông tin ngữ cảnh hoặc dữ liệu bổ trợ để điều chỉnh tham số lọc linh hoạt hơn. Chuỗi vận tốc GNSS sau tiền xử lý cũng có thể được khai thác hiệu quả cho các bài toán phân cụm trạng thái giao thông và phát hiện tắc nghẽn trong đô thị.

## TÀI LIỆU THAM KHẢO

[1]. Đỗ Văn Mạnh, Trần Quang Học, Lê Khánh Giang, Vương Xuân Cần, Vũ Văn Trường, Enhanced Deep Neural Networks for Traffic Speed Forecasting Regarding Sustainable Traffic Management Using Probe Data from Registered Transport Vehicles on Multilane Roads, Sustainability, 16 (2024) 2453. https://doi.org/10.3390/su16062453.

[2]. Đỗ Văn Mạnh, Đinh Tuấn Hải, Development of a Sustainable National Traffic Information Notification System: A GNSS-Based with Enhanced-LSTM for Urban Road Traffic Speed Forecasting, International Journal of Intelligent Transportation Systems Research, 23 (2025) 489-502. https://doi.org/10.1007/s13177-025-00463-2

[3]. Asakura Y, Hato E, Tracking survey for individual travel behaviour using mobile communication instruments, Transportation Research Part C: Emerging Technologies, 12 (2004) 273-91. https://doi.org/10.1016/j.trc.2004.07.010

[4]. D'Andrea, E. and F.J.E.S.w.A. Marcelloni, Detection of traffic congestion and incidents from GPS trace analysis, Expert Systems with Applications, 73 (2017) 43-56. https://doi.org/10.1016/j.eswa.2016.12.018

[5]. Ruwisch, F. and S.J.I.T.o.I.T.S. Schön, Feature Map Aided Robust High Precision GNSS Positioning in Harsh Urban Environments, IEEE Transactions on Intelligent Transportation Systems, 26 (2025) 13721 - 13733. https://doi.org/10.1109/TITS.2025.3569975

[6]. Ochieng, W. and K. Sauer, Urban road transport navigation: performance of the global positioning system after selective availability, Transportation Research Part C: Emerging Technologies, 10 (2022) 171-187. https://doi.org/10.1016/S0968-090X(02)00008-6

[7]. Juan C. Herrera, Danie B. Work, Ryan Herring, Xuegang (Jeff) Ban, Quinn Jacobson, Alexandre M. Bayen, Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment, Transportation Research Part C: Emerging Technologies, 18 (2010) 568-583. https://doi.org/10.1016/j.trc.2009.10.006

[8]. Davies, L. and U.J.J.o.t.A.S.A. Gather, The identification of multiple outliers, Journal of the American Statistical Association, 88 (1993) 782-792. https://doi.org/10.1080/01621459.1993.10476339

[9]. Hampel, F.R.J.J.o.t.a.s.a., The influence curve and its role in robust estimation, Journal of the American Statistical Association, 69 (1974) 383-393. https://doi.org/10.1080/01621459.1974.10482962

[10]. Hyndman, R.J. and G. Athanasopoulos, Forecasting: principles and practice, OTexts, ISBN 978-0-9875071-1-2. http://OTexts.com/fpp2/, 2018, May- 8.

[11]. Barrios, C., Y.J.I.T.o.I. Motai, and Measurement, Improving estimation of vehicle's trajectory using the latest global positioning system with Kalman filtering, IEEE Transactions on Instrumentation and Measurement, 60 (2011) 3747-3755. https://doi.org/10.1109/TIM.2011.2147670

[12]. Kalman, R.E., A new approach to linear filtering and prediction problems, Journal of Fluids Engineering, 82 (1960) 35-45. https://doi.org/10.1115/1.3662552