GNSS-Based Solutions Testing in an ERTMS Context: A Framework for Statistical Performance Analysis

Quentin Mayolle, Juliette Marais, Martin Fasquelle, Vincent Tardif, Emilie Chéneau-Grehalle

GNSS-Based Solutions Testing in an ERTMS Context: A Framework for Statistical Performance Analysis

1st Quentin Mayolle
IRT Railenium
Valenciennes, France
quentin.mayolle@railenium.eu

2nd Juliette Marais
Université Gustave Eiffel
Villeneuve d'Ascq, France
Juliette.marais@univ-eiffel.fr

3nd Martin Fasquelle
IRT Railenium
Valenciennes, France
martin.fasquelle@railenium.eu

4nd Vincent Tardif
Centre d'Ingénierie du Matériel - CIM SBF
SNCF
Le Mans, France
vincent.tardif@sncf.fr

5nd Emilie Chéneau-Grehalle
Centre d'Ingénierie du Matériel - CIM SBF
SNCF
Le Mans, France
emilie.cheneau@sncf.fr

***Abstract*—This article proposes a new methodology to label GNSS data from multiple railway environments, based on real measurements and fusion with external sources of information (satellite and infrared). A precise attribution of each GNSS observation to one specific environment is made possible, allowing the study of short time intervals (several seconds), and the analysis of local errors, computed with the Code-Minus-Carrier method. In addition, a complete probabilistic Bayesian model is employed to characterize the errors of each environment, based on stochastic processes to consider temporal correlations between errors. This new model is then analyzed and evaluated from the acquired data, and has the ability to generate new samples for simulation purposes. Finally, a Hidden Markov Model exploits the fitted model to perform a simple detection of the environment based on the local errors calculated.**

***Index Terms*—GNSS, railway applications, multipath, Bayesian inference, statistical modeling.**

I. Introduction

Europe identified GNSS for safety critical railway applications and to be integrated in rail in general as part of the toolset to help railway to contribute to reduce transport carbon footprint. Indeed, to increase the use of trains in European transports, railways must improve their attractiveness for passengers and freight, but also increase reliability, availability and efficiency by reducing capital expenditure and operational costs. GNSS is part of the global digitalization scheme of freight that aims to offer added value to the client's knowledge of accurate time of arrival and continuous monitoring of transport conditions. A major challenge will be to reach stringent applications. GNSS is today seen as a realistic and serious game changer for the future of the ERTMS (European Rail Traffic Management System). The localisation function is today performed with both odometry and balises. Odometer provides a continuous train position in time from a reference point. But as the distance delivered by the odometer shows a growing bias with distance, due to wear and wheel sliding, the use of on-track balises allows to reset this error. Future systems under development now will be based on on-board localisation solutions with GNSS receivers as part of a multi-sensor solution. It will allow the development of new concepts such as moving blocks, virtual coupling and automation. However, the environmental conditions of tracks and surroundings configuration, i.e, tunnels, dense urban areas or vegetation often degrade positioning performance and thus efficiency and safety. Many progresses have been made in the past years to develop more robust receivers, multi-sensor solutions (CLUG project [1]) or missing tools such as Digital Maps and some demonstrators are planned for 2025. But these developments must be accompanied by the development of assessment tools for demonstration plans and future certification. How can we evaluate performances in a dynamic environment (train, satellite, obstacles)? How can we be sure that every configuration has been tested? What is the impact of a failure (inaccuracy, missed detection) on operation? Some of these issues are addressed in the ongoing R2DATO project funded by Europe's rail and addressing the development of technologies in several fields of digital automated up to autonomous train operations, seeking a new paradigm in how the rail system is operated, increasing safety, flexibility, capacity, performance and reducing energy consumption and costs. As the performance of satellite-based positioning solutions varies over time and space, it is not possible to exhaustively demonstrate the performance of an on-board solution through long and costly test campaigns. Instead, this variety of scenarios can be carried out on a test bench, equipped with tools for simulating realistic signal reception conditions and sensor errors. However, today, no realistic

This work has been funded by FP2 R2Dato project funded under Europe's Rail.

models of these errors exist for railway environments. The challenge here is then to model them considering variations of the track surroundings and satellite positions in time.

GNSS errors are classically divided into global and local errors. Commercial signal simulators use to simulate global errors created by propagation through the atmosphere and by the system itself (orbit, clock, etc.). them. Local errors are, by definition, closely linked to the propagation environment close to the receiver and its antenna. They are therefore more difficult to model and simulate. In the past Gate4Rail project [2], initial error models have been specifically defined for the railway environment based on measurement campaigns. A few representative environments have been proposed (open sky, urban, forest), as well as the crossing of a few special features such as bridges and tunnels. However, these models are limited. They represent only partially the conditions encountered and are variable over time.

This paper proposes a date-driven framework for the generation of realistic railway local errors, based on real measurements from in-field experiments. The possibility to establish complex statistical models from multiple environments is a key aspect for future simulation tools to test and assess GNSS systems. The main objective of this paper is to provide a set of local errors models, each related to a specific type of environment: open-sky, vegetation, buildings,... with integration of the temporal dynamic of the errors in the context of train related GNSS acquisitions. The framework is composed of several steps.

First, we introduce a methodology to identify offline the environments crossed by a train during its journey. This spatial segmentation is based on a multi-source information fusion. A precisely timestamped ground-truth with estimated positions of the train is combined with public satellite and infrared images to perform easily a hand-made attribution between environments and observations. Detection of buildings and specific vegetation areas is performed to attribute one single environment to each individual GNSS observation. Additional relevant information such as the map elevation is included to localize specific perturbations: tunnels, bridges...

The second step is the inference of a Bayesian probabilistic model for local errors, involving analytical computations of the Code-Minus-Carrier (CMC) errors. Independent Gaussian errors are the preferred solutions in most of the GNSS statistical analyses, with modifications to increase the robustness of estimations. Here, we include the temporal dependences in the sequences of errors with stochastic process models. Non-Gaussian distributions, such as Laplace or Student's t distributions are tested to incorporate the realistic behaviour of the real errors. Very unlikely observations are therefore not removed and directly modelled. Their generations in future simulations provide challenging scenarios to test GNSS systems. Error models are based on the segmentation of the data into environments identified in the previous step. Dependences between the errors and additional available GNSS features, such as the elevation of satellites for instance, are also considered. A herarchical Bayesian framework allows a complete quantification of the full posterior distribution of the model parameters and uncertainty bound estimations, to be further integrated in a simulation process. The inference task is performed with modern Monter Carlo Markov Chain (MCMC) techniques, which can process a large quantity of data. Parameters estimated of each environment are compared to the open-sky situation to quantify the changes in dynamic properties. Notably, the increase of the errors amplitude in the building environment is reported.

The third and last step focuses on an auxiliary task induced by the establishment of the probabilistic models. The characterization of environments from the local errors data is included in a detection process to assess the predictive power of the inferred models. Based on the individual statistical distributions related to each environment, the most probable sequence of unobserved environments can be derived from any new trajectory (as soon as the GNSS observations are provided). This specific step provides practical results for comparison of the models. The conducted work is based on acquisitions campaigns, for which the ground-truth trajectory has been estimated. The data comes from real passenger trains and gathers hours of measurements in multiple environments.

Future work will include the validation of the framework based on the full simulation chain, relying on mixed hardware-software processes.

II. Multipath Statistical Modeling

A. Existing approaches for local error analysis

Trains pass through a wide variety of environments during their journeys. A train can easily move from one environment to another at high speed, spending only a few seconds in any one environment.

In GNSS applications, the problem of local errors, such as multipath, can hamper good localization of the train as it travels. Several solutions have been developed in the past to reduce the importance of local errors, through mathematical modeling of multipath [3], or the addition of 3-D building simulation methods [4].

Other solutions involve identifying the presence of multipath [5], or the presence of specific environment in which the train is located [6], and adapting mitigation solutions accordingly. Numerous techniques have been developed for this task, most recently using machine learning techniques based on labeled data [7], [8].

However, these approaches consist in decision tools. They don't allow a modeling of the errors, an important step to simulate errors at a later stage. Some characterization have been performed, such as the statistical description of multipath [9], with the definition of Gaussian bounds to be calculated to represent the train track. Statistical quantile based models, integrating multiples features, such as the Carrier-to-Noise ratio (C/N_0), the elevation or the number of tracked signals have also been developed to provide improved information about the errors [10].

Nevertheless, the statistical underlying models are often Gaussian distributions, which does not necessarily allow the

modeling of extreme event, which can append in specific environments. More advanced distribution with heavy tails have been suggested for GNSS applications [11]. Lastly, the models listed above ignore the temporal evolution of errors.

B. Environment labeling

A specific train trajectory is kept in this article, extracted from the datasets of the CLUG project. It contains multiple environments, to be described later. The acquisitions have been recorded from 16:40:00 to 23:59:44, starting at Fenouillet and ending at Foix (France), on April 3th 2021.

The CLUG datasets don't provide any label to the data (type of environment encountered by the train during the trajectory). The simplest solution found was to manually create these labels, using external sources of information about the type of environment. To perform this task, a human operator needed to analyses the location of the train during the trajectory, and associate to each timestamp one environment. Since the CLUG datasets easily exceed thousands of observations for one hour of measurements (classically, the GNSS observations are sampled at 1 Hz), labeling of each individual observation is time consuming, and not necessarily feasible.

The proposed solution is to segment the train trajectory manually into intervals of several thousands of meters. Next, for each GNSS observation, based on the estimated position of the train, a search of the nearest interval is performed. This process allows a fast attribution of an environment to each GNSS observation. In the GLUG datasets, ground-truth files are available, based on the information fusion between GNSS and IMU data. The position estimations from these files are therefore used for more precise attributions of an environment label.

A central aspect is the decision of the type of environment at a specific location of the train. The path is precisely known thanks to existing train lines maps. The main information to be integrated by a human operator for the labeling is the presence of trees or buildings besides the train track. An access to this data is possible in France trough the national platform Geoportail [12]. Two maps unveil the most relevant information:

- Satellite images: to identify building, city areas, water;
- Infrared images: this solution is particularly efficient for the detection of vegetation area, and segment the trees around the train track.

The illustration of the joint use of the two sources of information is represented in Fig. 1. The infrared view facilitates the identification of trees, and the satellite view allow the detection of buildings for a human operator.

Based on the previous observations, primary environment classes are defined, when the track is surrounded by a homogeneous environment (both sides of the tracks share the same environmental characteristics):

- Trees;
- Buildings;
- Open-sky;

Fig. 1. Infrared (left) and satellite (right) view from Geoportail of a Mixed-Trees-Buildings environment.

- Bridge;
- Post-bridge (the area coming just after the bridge);
- Train-station;
- Classification yard (specific area at the beginning and end of the train trip).

Besides these classes, most of the time the two sides of the tracks do note share the same properties: for instance one side is full of trees, and the other one is totally empty. In this situation, we define secondary classes, which are mixing of the previous ones:

- Mixed trees and open-sky;
- Mixed trees and buildings;
- Mixed buildings and open-sky;
- Open-sky in city (possible presence of buildings in the distance);

The full trajectory is represented in Fig. 2, with information about the main primary environment classes.

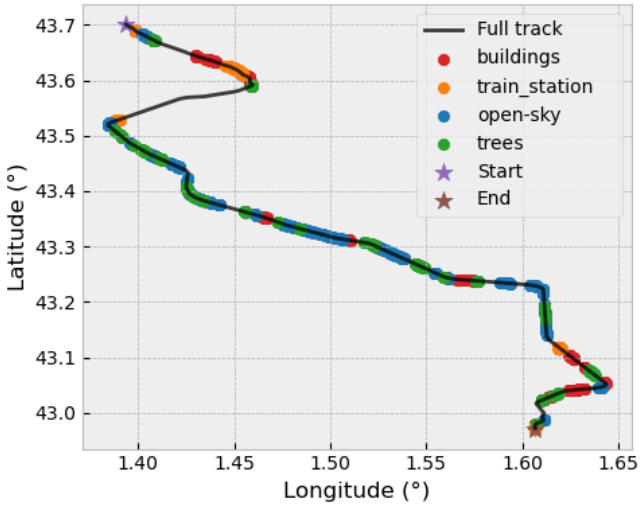

Fig. 2. Trajectory of the train with distributions of several environments: Buildings, Opens-Sky, Trees and Train-station.

Table I gathers the information related to the identified environments: their numbers, and statistics about their lengths (number of observations at 1 Hz included in the GNSS data).

TABLE I
STATISTICS ON NUMBERS OF OBSERVATIONS PER INTERVAL

Class	Count	Min	Mean	Max
Buildings	16	4	32.3	158
Open-sky	43	3	15.5	44
Trees	31	2	8.9	36

The means of the numbers of observation per environment never exceed 40 observations. This indicates that the train spends on average a short time in each a specific environment. The longest intervals of observations labeled with the same environments are found in the cities, with the environment "Buildings". On the contrary, the "Trees" environment interval are never really long. Indeed, many areas only have one side with trees.

As a results of the process, a total of 7194 GNSS observations have their own environment label. Among them, 665 are related to the "Opens-Sky", 517 to the "Buildings" and "275" to the "Trees".

C. Bayesian Inference

In parametric statistics, a distribution is represented by a parameter vector θ which gathers its characteristics. This vector contains all the required information to compute statistics (mean, variance, ...) and to simulate new samples from the distribution of interest. As an example, a Gaussian distribution is fully characterized by its mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}^+$. The notation for a new sample y is therefore $y \sim \mathcal{N}\left(\mu, \sigma^2\right)$. The parameter vector becomes $\theta = \left[\mu, \sigma^2\right]$. Simulation of samples for a distribution requires the knowledge of the vector θ. When this one is not identified, it must be derived from data (coming from sensors, experiments, ...). The estimation of the parameter θ is called the inference. In statistical analysis, maximizing a quantity that is representative of the fitness of a parameter according to data samples is usually performed based on the explicit calculation of the data likelihood. With notation D to represent the data sample set, the data likelihood with respect to a parameter vector θ is noted $\mathcal{L}(\theta) = p\left(D|\theta\right)$, with the right term describing the probability of the data given the parameter vector.

Inference can simply be done with the search of the parameter vector that maximizes the data likelihood. This method is the Maximum Likelihood Estimation (MLE):

$$\theta^{\text{MLE}} = \arg\max_{\theta} p\left(D|\theta\right) \tag{1}$$

However, this approach does not use any previous information about θ, and cannot include specific constraints about the possible values (for instance, restrictions on an interval). In the context of Bayesian inference [13], the focus is made on the probability distribution of the parameter θ given the data: $p\left(\theta|\, D\right)$, which has easier interpretation, and allows the calculation of the probability that the parameter belongs to an interval, for instance. Bayes' rules is used to calculate this distribution:

$$p\left(\theta|\, D\right) = \frac{p\left(\theta|\, D\right)p(\theta)}{p(D)} \tag{2}$$

The term $p(\theta)$ is called the prior. It represents the initial information we have about the parameter. It can either be informative or not. The term $p(D)$ is the data evidence, and represents the data fitness given the model (assumed distribution). It is complex to estimate, however its direct calculation is not necessary in most Bayesian inference techniques. The parameter vector that maximizes the probability $p\left(\theta|D\right)$ is called the Maximum A Posteriori (MAP):

$$\theta^{\text{MAP}} = \arg\max_{\theta} p\left(\theta|D\right) \tag{3}$$

However, a full understanding of the parameter distribution is sometimes needed, since the vector estimations θ^{MLE} and θ^{MAP} provide no information about the uncertainty related to the parameter. The analytic calculation of $p\left(\theta|D\right)$ is generally impossible, and approximation techniques are classically employed in Bayesian inference. A common technique is to estimate samples from the distribution of interest, using advanced techniques such as MCMC. This task is now easily done with the recent development of Probabilistic Programming Languages (PPL), such as PyMC [14] used in this article for the MCMC sampling. The only task needed is the definition of the distributions used, and the priors assumed on the parameter vector of interest.

The dataset D is now assumed to be composed of N observations, with $D = \{y_1, ..., y_N\}$. In Bayesian inference, the fitness of a model to the dataset D is computed with the Expected Log Pointwise Predictive Density (eldp_{LOO}) defined as [15]:

$$\text{eldp}_{\text{LOO}} = \sum_{i=1}^{N} \log p(y_i \mid y_{-i}) \tag{4}$$

where $p(y_i \mid y_{-i})$ is the leave-one-out predictive density of the model given the data without the ith data point, defined as

$$p(y_i \mid y_{-i}) = \int p(y_i \mid \theta) p(\theta \; y_{-i}) d\theta \tag{5}$$

Fortunately, in Bayesian inference with samples from the posterior distribution, this quantity is easily estimated with a procedure called Importance Sampling, which prevent the repetition (N times) of the inference process for each observation of the dataset. By comparing different models with their respective eldp_{LOO}, the practitioner selects the model with the higher score. A difference of eldp_{LOO} of 1 is considered as low (no model significantly better than the other), whereas a difference above 4 is considered as strong. In the following, the observations y_i to be studied will be the local errors.

D. Multipath Calculation

The Code-Minus-Carrier (CMC) method is employed to estimate local errors in this paper. The calculation if based on the following GNSS measurements, the pseudorange R_i and the carrier phase Φ_i, for any frequency-band $i \in \{1, 2, 5\}$, that are decomposed as:

$$R_i = \rho + c\left(\delta t_{\text{rcv}} - \delta t^{\text{sat}}\right) + Tr + I_i + M_i + \varepsilon_i \quad (6)$$

$$\Phi_i = \rho + c\left(\delta t_{\text{rcv}} - \delta t^{\text{sat}}\right) + Tr - I_i + N_i\lambda_i + m_i + \epsilon_i \quad (7)$$

With:

- ρ: geometric range;
- Tr: tropospheric delay;
- I_i: ionospheric delay for frequency band i;
- N_i: integer ambiguity;
- M_i: multipath;
- m_i: carrier phase multipath;
- ε_i: pseudorange error
- ϵ_i: carrier phase error.

It is generally assumed that $m_i \ll M_i$ and $\epsilon_i \ll \varepsilon_i$

In the following, the multipath M_i and the error ε_i terms are gathered to produce the local error term e_i such that:

$$e_i = M_i + \varepsilon_i \quad (8)$$

The CMC method expresses the following quantity:

$$\begin{aligned} \text{CMC}_i &= R_i - \Phi_i && (9)\\ &= 2I_i + M_i - m_i - N_i\lambda_i + \varepsilon_i - \epsilon_i && (10)\\ &\approx 2I_i + M_i - N_i\lambda_i + \varepsilon_i && (11)\end{aligned}$$

The link with the local error appears then from the CMC quantity:

$$e_i \approx \text{CMC}_i - 2I_i + N_i\lambda_i \quad (12)$$

The main objective is to model the error e_i, which is assumed to be stochastic, and follow a parametric probability distribution. Two different hypothesis are investigated:

1) Independent errors: errors are assumed independent between them. There is no temporal correlation in the series of errors.
2) Temporal correlation: to allow more complex modeling, correlations inf the series of errors can be modeled.

The estimation of the ionospheric delay can be precisely made with the carrier-phase, and while therefore contain the ambiguity component, noted $\tilde{I}_i$. Using two frequency band i and j, with respective frequencies f_i and f_j,

$$\tilde{I}_i = f_j^2 \frac{\Phi_i - \Phi_j}{f_i^2 - f_j^2} \quad (13)$$

The difference $\text{CMC}_i - 2\tilde{I}_i$ still contains an ambiguity term, which should be removed. The mean of the signal can be removed since this ambiguity term does not change if no cycle-slip occur. The last step is to identify the possible jumps due to cycle-slips. A multi-frequency cycle-Slip detectors is implemented to perform this task (two frequencies used), using the geometry-free combination $\Phi_i - \Phi_j$. Since the data processing is here performed offline, the jumps can be detected, and the mean value subtracted from the whole signal between them. A possible alternative is the Hatch filter, which can be applied sequentially (online use), but requires more parameters to be selected. In this article, the mean value is subtracted. Only GNSS observations related to the GPS constellations are analyzed, but the methodology can be generalized to any constellation.

III. Bayesian modeling

A. Single Satellite Statistical Analysis

A first preliminary analysis involves only the satellite with the most important number of observations : GPS 27. The estimated CMC errors are represented in Fig. 3. Some extreme values are visible around points 1000 and 1500. The aim of this section is to illustrate the poorness of Gaussian fitting, and suggest different distributions to model the errors, which could later be use for simulation purposes. The complexity of the latter complex distribution is nearly equivalent to the one of the Gaussian distribution.

An introduction of 3 major distributions is made here. The Gaussian distribution, whose parameters are the mean μ and the variance σ^2 , with probability density $p(x|\mu, \sigma^2)$ as:

$$\begin{aligned} p(x|\mu,\sigma^2) &= \text{Normal}(x|\mu,\sigma^2) && (14)\\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) && (15)\end{aligned}$$

The Laplace distribution with localization parameter m and spread parameter b, with probability density:

$$\begin{aligned} p(x|m,b) &= \text{Laplace}(x|m,b) && (16)\\ &= \frac{1}{2b} \exp\left(-\frac{|x-m|}{b}\right) && (17)\end{aligned}$$

The Student's t distribution include a third parameter, called the degree of freedom ν besides the location parameter m and the spread parameter s. The Gaussian distribution is a limit of the Student's t distribution when $\nu \to +\infty$. Its probability density is:

$$\begin{aligned} p(x|m,s) &= \text{StudentT}(x|m,s) && (18)\\ &= \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{s\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{(x-m)^2}{\nu s^2}\right)^{-\frac{\nu+1}{2}} && (19)\end{aligned}$$

Both the Laplace and the Student's t distributions have heaver tails than the Gaussian distribution. There are less sensitive to outliers values, and can better model events which are far from the mean. Here, we expect outliers values (for instance extreme multipath due to local object such as buildings) to happen.

The data likelihood of the three fitted distribution, with MLE method, is gathered for frequency bands L1 and L2 in table

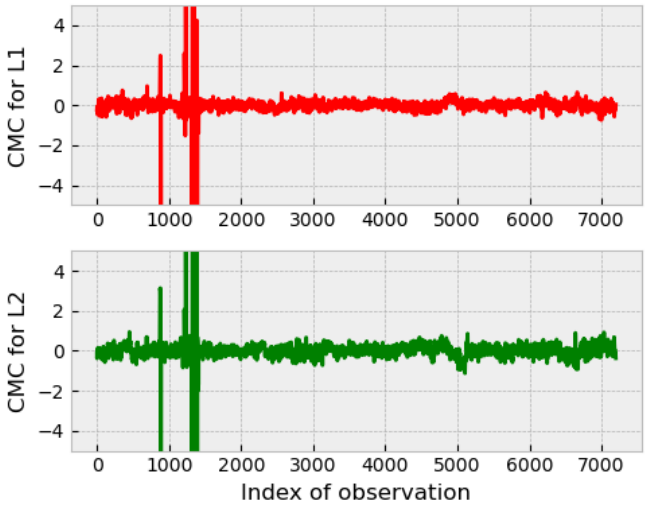

Fig. 3. CMC calculated for satellite G27 (frequency bands L1 and L2).

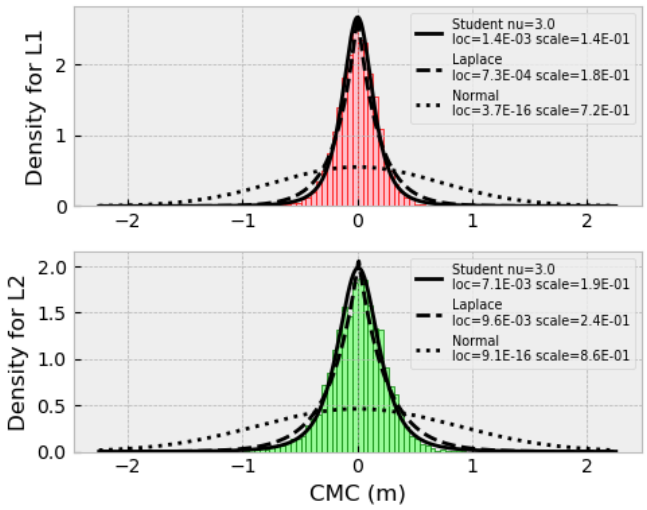

Fig. 4. Estimated parametric CMC distributions for satellite G27.

II. The visual fitting is provided in Fig. 4. A comparison of likelihood with the base method, the Gaussian distribution, is provided. The comparison if made with the factor between the distribution studied and the Gaussian one. Since this ratio can take very low or high values, the logarithm of this ratio is taken. If the log ratio is superior to zero, this indicates that the distribution better fits the data than the Gaussian distribution. Finally, in the given table, the two distributions, Student's t and Laplace, always outperform the Gaussian. The Student's t distribution is still superior to the Laplace one.

The conclusions are:

- The Gaussian distribution fit poorly the data, visually there is a huge gap between the fitted density and the histograms. Extreme values have a great impact that causes an over-dispersion of the distribution. The Laplace and Student's t distributions fit nearly identically the histogram (visually).
- The L2 band shows higher dispersion (higher scale parameters) than the L1 band. This highlights the importance of independent models for each frequency band.
- The Student's t distribution provides the best data likelihood. The extra parameter ν allow the distribution the control the importance of outlier values, and give therefore more flexibility to the model.

TABLE II
LOG OF THE RATIO OF LIKELIHOODS WITH GAUSSIAN FITTED MODELS

Distribution	L1 Band	L2 Band
Student' t	9160	8350
Laplace	7660	7020

In the following, only the L1 frequency band will be kept, and all CMC errors will be related to this band. The CMC values of all GPS satellites are gathered in the dataset D. Consequently, based on the GNSS observations available and the visible satellite, the total number of CMC errors to be processed equals 47872 (then N = 47872).

B. Time Series Analysis

The next step is the analysis of temporal dependencies inside the CMC errors calculated. An auto-regressive process x of order p is defined, using p scalars $a_1, ..., a_p$ as

$$x_t = a_1 x_{t-1} + ... + a_p x_{t-p} + \epsilon_t \tag{20}$$

With ϵ_t a sample from a distribution, from instance a Gaussian distribution with mean μ and variance σ^2: $\mathcal{N}\left(\mu, \sigma^2\right)$. More advanced distributions can be used to adapt different perturbations, such as the Student's t distribution, which offer better modeling of extreme values [16]. This process introduces a dependence between successive samples. The level of correlation between samples is managed by the order of the process (the higher the order, the longer the dependence will occur).

An auto-regressive process will insert a memory effect in the data, which could be used to model local spatial event which share same properties, such as a group of buildings, or trees of same heights in a forest.

Given a specific time series, the objective of the modeling process is to find the order which best corresponds to the signal. For this, visual tools, such as the Auto-correlation and Partial Autocorrelation graphs are generally employed. A quantitative method to find the optimal order of a process is to fit multiples AR process with different order, and search the process that better fit the time series according to a specific criteria. In time series analysis, such criteria is either the Aikaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), which should be minimized (the less the value, the better fit).

The auto-regressive model order is selected with the shape of the AIC and BIC curves. Multiple rules exist for the

selection of the order: value that minimizes the AIC or BIC, value of the curve inflections or values where a plateau is reached. AIC and BIC curves obtained from the GPS satellites, in Fig. 5, exhibit a strong decrease between orders 1 and 4, and then smoothly converge. A maximum order of 5 will then be considered.

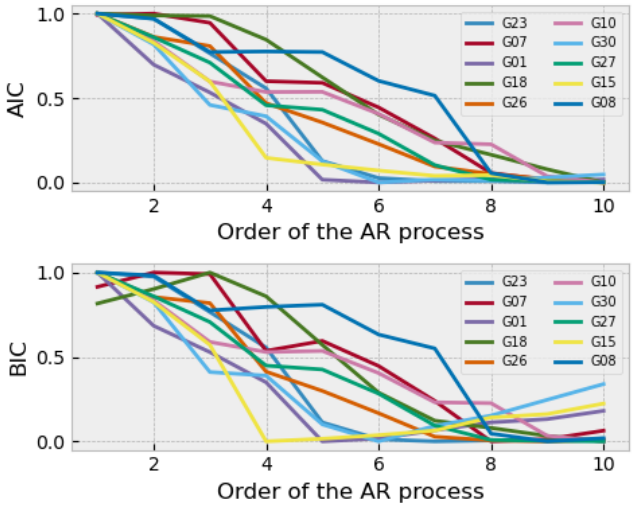

Fig. 5. AIC and BIC curves for satellite G27.

C. Choice of the Bayesian Models

Three major models are now introduced to study the statistical characteristics of local errors for L1 frequency band with the GPS constellation. The observation of the models are the estimated multipath errors, noted as CMC_i for an error generated by the ith environment (with $0 \leq i \leq 10$ for our 11 environments).

1) Model 1: independent Gaussian observations: A first simple model is to assume independent observations, sampled from a Normal distribution, with one specific standard-deviation parameter $\sigma_{0,i}$ for each environment i. A weakly informative prior is selected (Half-Cauchy):

$$\sigma_{0,i} \sim \text{HalfCauchy}(1) \tag{21}$$
$$\text{CMC}_i \sim \text{Normal}(\mu = 0, \sigma = \sigma_{0,i}) \tag{22}$$

For all other models, a Student's t distribution, more robust to outliers and therefore more reliable, is chosen. The degree of freedom of this distribution, ν is unknown. The inference on this parameter is therefore also performed, using a weakly informative prior: the Inverse Gamma distribution with mean and scale parameters 1 and , noted $\text{InvGamma}(1, 1)$. Each environment is assumed to have its own degree of freedom ν_i.

2) Model 2: independent observations, variance depends on elevation (polynomial): The third model introduces a generic form for the standard-deviation. Variance models developed for the pseudorange measurements usually involve either a non-linear relationship with the observation C/N_0, or an inverse relationship with the sine of the elevation [17]. To provide more flexibility, a general polynomial of the elevation is selected. Contrary to the previous presented models, it does not strictly impose a decrease of the local error variance with the elevation, allowing a flexible adaptation to the relationship between the two variables, and inspired by previous existing models [18], [19]. Again, the parameter $\sigma_{0,i}$ correspond to the value of standard deviation for the maximal elevation. It will be called the initial standard deviation. However, in this specific situation, no specific trend is imposed to the model. The polynomial has two parameters : $\beta \in \mathbb{R}$ and $\gamma \in \mathbb{R}$. Only two parameters are needed here, since the constant part of the polynomial is implicitly gathered in the parameter $\sigma_{0,i}$.

$$\sigma_{0,i} \sim \text{HalfCauchy}(1) \tag{23}$$
$$\nu_i \sim \text{InvGamma}(1, 1) \tag{24}$$
$$\beta, \gamma \sim \text{Normal}(0, 1) \tag{25}$$
$$\sigma_i = \sigma_{0,i}\, e^{\beta(E-\pi/2)}\, e^{\gamma(E-\pi/2)^2} \tag{26}$$
$$\text{CMC}_i \sim \text{StudentT}\,(\mu = 0, \sigma = \sigma_i, \nu) \tag{27}$$

3) Model 3: hierarchical Auto-regressive: The last model introduces the auto-regressive modeling. For a specific CMC value at time t, its previous values at times $t-1$ up to $t-5$ are gathered inside the vector x. The vector α of dimension 5 gathers the auto-regressive coefficients to be estimated.

A higher level of Bayesian modeling can be achieved by considering a hierarchical model [13], which offers a form of regularization of the model. The parameters related to each environment is assumed to come from a population distribution. This distribution is here related to one other prior distribution, called a hyperprior. This hyperprior is here the Exponential distribution, related to the beta parameter β_{Cauchy} of the Half-Cauchy distribution, prior of the standard-deviation parameter $\sigma_{0,i}$:

$$\beta_{\text{Cauchy}} \sim \text{Exponential}(1) \tag{28}$$
$$\sigma_{0,i} \sim \text{HalfCauchy}(\beta_{\text{Cauchy}}) \tag{29}$$
$$\nu_i \sim \text{InvGamma}(1, 1) \tag{30}$$
$$\beta, \gamma \sim \text{Normal}(0, 1) \tag{31}$$
$$\alpha \sim \text{Normal}(0, 10) \tag{32}$$
$$\sigma_i = \sigma_{0,i}\, e^{\beta(E-\pi/2)}\, e^{\gamma(E-\pi/2)^2} \tag{33}$$
$$\text{CMC}_i \sim \text{StudentT}\left(\mu = x^t\alpha, \sigma = \sigma_i, \nu = \nu\right) \tag{34}$$

The hierarchical model introduces a dependence between the different environment initial standard deviations. This constraint prevents them from having enormous values. One important consequence of the model is also the possibility to simulate later an external type of environment, by sampling a value from the posterior $p(\sigma_{0,i} \mid D)$. For instance, if an environment is not clearly identified (by an operator or a software), a simple solution is to take the mean value of this

posterior (mean standard-deviation of the classes), and use this mean to produce samples of local errors.

IV. Results

A. Model comparison

Table III gathers the scores of the different models, with their ranking, and the difference of scores with the best model, which is the model 3. The auto-regressive model has clearly the best scores, but the gap between the second best model (Model 2) and the Gaussian model (Model 3) is even higher. This table clearly indicate that the Gaussian assumption is a poor choice for statistical analyses. The auto-regressive assumption is validated by this experiment. The next step in the analysis of the posterior distributions estimated by the MCMC sampling, for the best model (Model 3).

TABLE III
Ranking of model based on their expected log pointwise predictive densities

Bayesian model	Rank	eldp	eldp difference
Model 3 (AR Student's t)	1	11553	0
Model 2 (Student's t)	2	-20187	31740
Model 1 (Gaussian)	3	-72471	84024

B. Analysis of auto-regressive parameters

Table IV gathers the statistics of the posterior distribution related to the auto-regressive parameters, estimated with the hierarchical model (Model 3). The mean of the posterior distributions for each coefficient is indicated with plus or minus the standard deviation of the posterior.

TABLE IV
Estimations of parameters from posterior distributions and least-square method

Coefficient	Posterior mean $\pm$ standard-deviation	Least-Square
a_1	0.952 $\pm$ 0.002	0.375
a_2	-0.027 $\pm$ 0.003	0.059
a_3	-0.006 $\pm$ 0.001	-0.021
a_4	-0.006 $\pm$ 0.002	-0.093
a_5	-0.009 $\pm$ 0.002	-0.091

The dependencies introduces by the choice of the environment and the elevation has a clear effect on the estimated auto-regressive model parameters. The weights of the coefficients a_2, a_3, a_4 and a_5 are reduced to favor the first one a_1. The temporal dependencies are reduced, and introduce more stochastic variability in the process. The intensity of the perturbations is controlled by external information, related to the location of the train (the environment), and the geometry of the estimation process (location of satellites).

C. Analysis of environment parameters

The following table V provides all specific estimated parameters related to environments. It provides the means of the posterior distributions with the estimated standard deviation. The initial standard deviation σ_0 parameter control the perturbation level introduced in the auto-regressive stochastic process, whereas the degree of freedom parameter ν models the heaviness of the distribution tail. An environment with a low degree of freedom exhibits more frequently extreme values of CMC (very high in absolute value). A general representation of posterior distributions and samples for all environment is included in Fig. 7 as illustration of the MCMC sampling process.

TABLE V
Estimations of environment parameters from posterior distributions

Environment	$\sigma_{0,i}$	ν
0 : mixed-tree-build	0.057 $\pm$ 0.002	2.486 $\pm$ 0.068
1 : classification yard	0.065 $\pm$ 0.002	5.601 $\pm$ 0.489
2 : post-bridge	0.079 $\pm$ 0.020	0.595 $\pm$ 0.105
3 : open-sky-urban	0.065 $\pm$ 0.003	3.439 $\pm$ 0.370
4 : mixed-build-open	0.057 $\pm$ 0.002	2.423 $\pm$ 0.121
5 : open-sky-rural	0.054 $\pm$ 0.002	2.544 $\pm$ 0.097
6 : train-station	0.057 $\pm$ 0.001	2.932 $\pm$ 0.085
7 : mixed-tree-open	0.053 $\pm$ 0.001	2.033 $\pm$ 0.049
8 : tree	0.054 $\pm$ 0.002	2.353 $\pm$ 0.148
9 : bridge	0.121 $\pm$ 0.032	0.737 $\pm$ 0.161
10 : building	0.056 $\pm$ 0.002	1.463 $\pm$ 0.053

Similarly, table VI provide the estimations from the posterior distribution of the polynomial of order 2 that models the dependence to the elevation. Fig. 6 illustrate the resulting standard deviations for "Buildings" and "Open-Sky" environments, with use of their initial standard deviations.

TABLE VI
Estimations of common parameters from posterior distributions

Environment	β	γ
All	0.039 $\pm$ 0.062	0.611 $\pm$ 0.034

Among the different environments, the "Buildings", "Bridges" and "Post-bridges" exhibit the lower degree of freedom ν_i, above 2. However, the initial standard deviation for the "Buildings" environment is tight to the ones of the "Trees" and "Open-Sky", as represented in Fig. 8. The presence of buildings has therefore a localized impact on the CMC: several extremely high values occur sometimes, giving to their distribution a heavy-tail. The case of bridges is the most severe, with both higher initial standard-deviation and a low degree of freedom (inferior to 1). CMC values in this areas are significant, and the probability of very high abnormal value is high, even if these environments are encountered for a very short period by the train.

The open-sky situation, as expected, gathers the lowest initial standard deviation with the higher degree of freedom. Few outliers are expected in such environment, and the estimated parameters comfort this expectation.

D. Environment detection with a Hidden Markov Model

An auxiliary task is now investigated. The objective is to construct a classification method to identify the environment

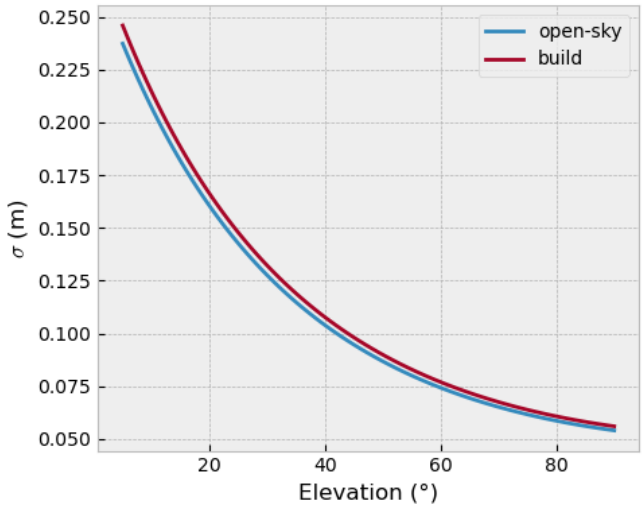

Fig. 6. standard deviations for "Buildings" and "Open-Sky" environments, depending on the elevation.

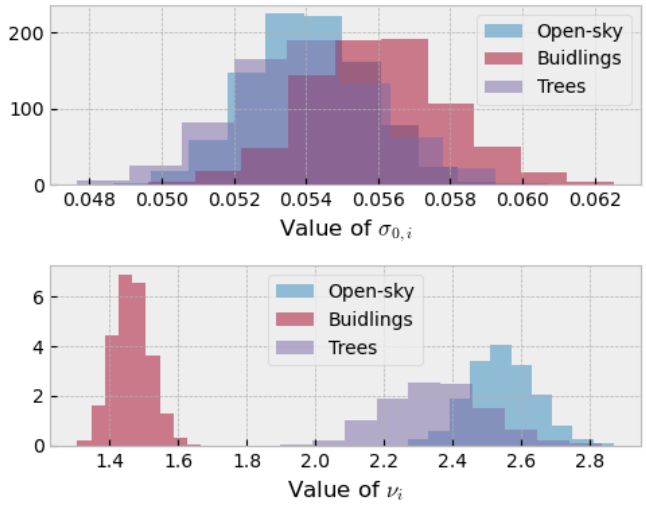

Fig. 8. Histograms of statistical scale parameter and degree of freedom for the main environments.

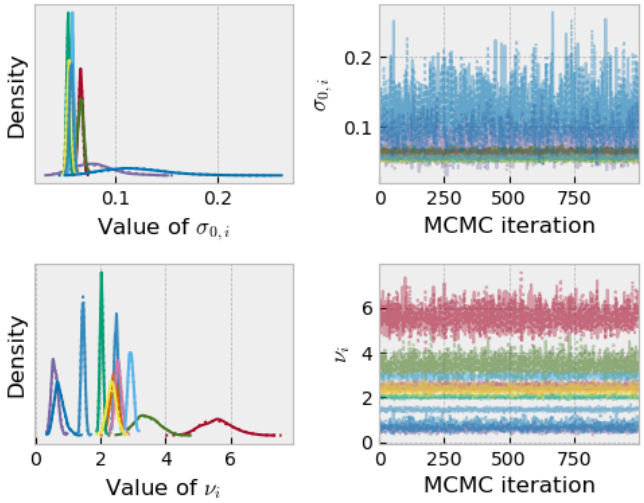

Fig. 7. Density and samples for initial scales and degrees of freedom for all environments.

based only on the CMC errors. Many models assume the independence between observations (for two observations s_i and s_j such that $i \neq j$, s_i and s_j). The order of observations has no importance for the data modeling process. Each observation is drawn from a distribution, specific to one environment. However, this rough hypothesis does not take into account the temporal dependencies between observations and the possible correlations between observations. For example, some events, like occlusion, disturb several points. A classical hypothesis to take into account this temporal dependencies is the Markov hypothesis : the current state only depends on the previous state. A statistical tool to study discrete stochastic process is the Hidden Markov Model (HMM), which can act as a detector, for instance to detect indoor and outdoor environments [20], or distinguish between more environments in our situation.

For one hidden state s_i at step i which belongs to the class $k \in \{1, ..., K\}$ among K classes, the transition is possible in the next step $i+1$ to all the classes with a probability. The probability to move from state k to state l is noted $\pi_{kl} \geq 0$.

A Hidden Markov Model is defined by the transition matrix P:

$$\begin{bmatrix} \pi_{11} & \pi_{12} & \dots & \pi_{1K} \\ \pi_{21} & \pi_{22} & \dots & \pi_{2K} \\ \vdots & & \ddots & \vdots \\ \pi_{K1} & \dots & \dots & \pi_{KK} \end{bmatrix}$$

The transition matrix P is called a transition matrix: all its coefficients are positive, and the sum of each row equals 1.

In the case of environment detection, the hidden state corresponds to an unknown environment in which the train is, that generates a local error. The statistical model for measurement generation is described by the Student-t distribution which uses the scaling factor integrating the dependence to the elevation. The HMM, given the statistical models of each environment and given the transition matrix, can estimate the most probable sequence of hidden states with the Viterbi algorithm.

The HMM is here employed as an environment detector, based only on the sequence of CMC errors calculated. To focus on classical environments commonly identified in the literature, only the 3 main primary environments: "Open-sky", "Buildings" and "Trees" (thus K=3) are retained. The objective is to assess the performances of the detector. For this purpose, for each epoch identified as one of the three later environments, the estimated environment is checked. The accuracy scores are defined as the ratio of epochs belonging to the three classes that are identified by the HMM as this environment over the total of epoch for this environment.

Since many environments are omitted, the real transition matrix cannot be estimated. For this purpose, a generic diagonal matrix is chosen: for $i \in \{1, 2, 3\}$ then $\pi_{ii} = 0.95$, and for $i \neq$ then $\pi_{ij} = 0.025$. This corresponds to a chance of changing the environment every 20 epochs, which is coherent with the lengths of observed intervals. The accuracy scores are gathered in the table VII.

TABLE VII
ACCURACY SCORES FOR ENVIRONMENT IDENTIFICATION WITH HMM

Satellite	Score Open-sky	Score Buildings	Score Trees	Count
G27	1.000	0.181	0.000	7027
G10	0.979	0.166	0.000	6960
G08	0.937	0.553	0.000	6848
G23	0.643	0.279	0.000	6702
G26	0.963	0.317	0.000	6056
G07	0.751	0.500	0.000	4327
G18	0.897	0.528	0.000	3589
G32	0.291	0.696	0.000	1774
G30	0.875	0.098	0.000	1756
G01	0.043	0.061	0.000	1671
G15	0.443	1.000	0.000	1162

The performances of detection are good for the "Open-sky" environment, degrade for the "Buidlings", and are null for the "Trees". This could be caused by the strong proximity of the estimated parameters between "Open-sky" and "Trees" environments. As a general conclusion, the HMM as interest when the CMC characteristics are significantly different. But a precise separation between environments solely based on CMC errors is not realistic.

V. CONCLUSION

The paper introduced several statistical tools to estimate the characteristics of local errors, based on prior calculation with the CMC method. The Bayesian paradigm allows a total quantification of the posterior distribution of model parameters with Monte Carlo Markov Chain sampling. A comparison between several models has been proposed, and the superiority of the more advanced model demonstrated. This model assumed a dependence between the elevation and the perturbation of the CMC error, and introduced temporal dependencies between consecutive CMC errors. The obtained model can later be used to generate new samples for testing purposes. Moreover, the performance of a statistical detector of environment based on the estimated parameters were assessed. However, the limitation of this method are serious, and were not able to distinguish environments which produce similar (albeit different) local errors.

REFERENCES

[1] https://www.clug2.eu/

[2] Gate4rail Deliverable 3.2 – Models for Fail-Safe positioning components w.r.t. Faults, 07/2020. https://gate4rail.eu

[3] Q.H. Phan, S.L. Tan, and I. McLoughlin, GPS multipath mitigation: a nonlinear regression approach. GPS Solutions 17, 2013, p371–380.

[4] L. T. Hsu, Y. Gu, Y. and S. Kamijo. 3D building model-based pedestrian positioning method using GPS/GLONASS/QZSS and its reliability calculation. GPS Solut 20, 413–428 (2016).

[5] L. -T. Hsu, GNSS multipath detection using a machine learning approach, 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, 2017, pp. 1-6.

[6] Y. Wang, P. Liu, Liu Q, M. Adeel, J. Qian, X. Jin and R. Ying, Urban environment recognition based on the GNSS signal characteristics. NAVIGATION. 2019; 66: 211–225.

[7] F. Feriol, Y. Watanabe, D. Vivet. GNSS-based environmental context detection for navigation. 2022 IEEE Intelligent Vehicles Symposium (IV), Jun 2022, Aachen, Germany. pp.888- 894.

[8] A. Siemuri, K. Selvan, H. Kuusniemi, P. Valisuo and M. S. Elmusrati, A Systematic Review of Machine Learning Techniques for GNSS Use Cases, IEEE Transactions on Aerospace and Electronic Systems, vol. 58, no. 6, pp. 5043-5077, 2022.

[9] A. Kliman, F. Roessl, A. Grosch, O. G. Crespillo, Railway GNSS Multipath Error Modelling Approach with both Train-Side and Operational Environment Characterization, Proceedings of the 37th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2024), Baltimore, Maryland, September 2024, pp. 114-126.

[10] F. Roessl, O. G. Crespillo, Robust GNSS Multipath Error Modeling Based on Deep Quantile Regression with Gaussian Overbounding, Proceedings of the 37th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2024), Baltimore, Maryland, September 2024, pp. 1402-1415.

[11] D. A. Hsu, Long-Tailed Distributions for Position Errors in Navigation. Journal of the Royal Statistical Society. Series C (Applied Statistics), vol. 28, no. 1, 1979, pp. 62–72.

[12] https://www.geoportail.gouv.fr/

[13] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin, Bayesian Data Analysis, 3rd Edition, CRC Press, 2013.

[14] O. Abril-Pla, V. Andreani, C. Carroll, L. Dong, C. J. Fonnesbeck, M. Kochurov, R. Kumar, J. Lao, C. C. Luhmann, O. A. Martin, M. Osthege, R. Vieira, T. Wiecki, R.Zinkov , PyMC: A Modern and Comprehensive Probabilistic Programming Framework in Python, 2023.

[15] A. Vehtari, A. Gelman, A.and J. Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, Stat Comput 27, 2017, pp 1413–1432.

[16] N. Balakrishna, Non-Gaussian Autoregressive-Type Time Series, Springer Singapore, 2022.

[17] B. Li, L. Lou, and Y. Shen, Gnss elevation-dependent stochastic modeling and its impacts on the statistic testing, Journal of Surveying Engineering, vol. 142, no. 2, 2015, p. 04015012.

[18] P. Teunissen and O. Montenbruck, Springer Handbook of Global Navigation Satellite Systems. vol. XXXI, Springer, 2017.

[19] T. Suzuki, Y. Nakano, Y. Amano, NLOS Multipath Detection by Using Machine Learning in Urban Environments, Proceedings of the 30th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GNSS+ 2017), Portland, Oregon, September 2017, pp. 3958-3967.

[20] H. Gao, and P. D. Groves, Environmental Context Detection for Adaptive Navigation using GNSS Measurements from a Smartphone. J Inst Navig, 2018, 65: 99–116.