A Tool for GNSS Integrity Verification Based on Statistical Extreme Value Theory

Henk Veerman, Arnaud van Kleef, Frank Wokke
NLR (National Aerospace Laboratory), The Netherlands

Bas Ober
Integricom, The Netherlands

Christian Tiberius, Sandra Verhagen
Delft University of Technology, The Netherlands

André Bos, Arjan Mieremet
Science & Technology Corp., The Netherlands

BIOGRAPHIES

Henk Veerman received his MSc in Experimental Physics at the University of Amsterdam. Now as a senior scientist at NLR, he works on novel technologies for application in aeronautics, with a focus on GNSS, data link communication and optical measurement techniques.

Arnaud van Kleef received his MSc. degree in Aerospace Engineering at Delft University of Technology. Arnaud participates in the development of a GNSS integrity monitoring tool for EGNOS and Galileo based on Extreme Value Theory. Furthermore, Arnaud is responsible for the operation and maintenance of NLR's Celeste GNSS monitoring facility for GPS, EGNOS, GLONASS and Galileo.

Frank Wokke is a senior R&D manager at NLR. Since 2002 he has been involved in the EGNOS and Galileo programmes and various other projects focusing on GNSS performance and validation.

Bas Ober received his PhD in 2003 from Delft University of Technology on GNSS integrity prediction and monitoring. He founded Integricom in 1998 and has continued to stay active in the field of GNSS performance assessment ever since.

Christian Tiberius is an associate professor at Delft University of Technology. He is involved in GNSS positioning and navigation research since 1991, currently with emphasis on data quality control, SBAS and Precise Point Positioning.

Sandra Verhagen Sandra Verhagen obtained her Ph.D. and M.Sc. from Delft University of Technology and is now an assistant professor at the same university. Her research focuses on ambiguity resolution and integrity monitoring for real-time kinematic GNSS applications, for which she received a research grant from the Dutch Science Organisation (NWO). She is president of Commission 4 Positioning and Applications of the International Association of Geodesy (IAG).

André Bos obtained a PhD in Computer Science from the Delft University of Technology, and co-founded Science & Technology Corp. He specializes in signal and data analysis, and is involved in the In Orbit Validation of the Galileo GNSS.

Arjan Mieremet received his PhD in Physics (Optics) at the Delft University of Technology in 2003. Since then Arjan has been research scientist, project leader and program leader in various fields varying from payload design, laser sensors (laser range finders, 3D-laser scanners, laser retro-reflection) and imagers (optical, infrared, radar and X-ray).

ABSTRACT

With the advent of augmentation systems such as WAAS and EGNOS, satellite navigation gets more and more used for so-called Safety of Life (SoL) applications, with civil aviation being the most prominent example. For SoL applications reliability performance of the system is typically expressed in terms of accuracy, availability, continuity of service and integrity. Especially the integrity of the position solution resulting from the satellite navigation signals is crucial. In general integrity performance requirements are exceedingly demanding, and therefore compliance is difficult to test. For example,

the GNSS integrity performance requirement for ICAO LPV (Localizer Performance with Vertical Guidance) approach procedures is 2 x 10^{-7} per approach position error probability of exceedance of the corresponding protection level, where an approach is defined as of 150 seconds duration. In particular, it is required to assess and verify the probability that the position error (PE) of a GNSS user exceeds the alert limit, while no alert is raised within the time-to-alert associated with the operation. Testing a GNSS system's integrity by collecting Misleading Information (MI) events not only is impractical given the low event rate - on average only one event is expected to occur every 20 years - but also the low yield is considered inadequate for a precise statistical assessment. It should be noted that although no MI events are observed the system can be unsafe, because of the fact that the probability of MI is higher than its integrity function, implemented through protection levels, promises: in this case we speak of *misleading integrity information* (MII). Thus MI is misleading *actual* position information, which is almost never present in the data, while we are looking for MII: for example SBAS PL's or Galileo's Integrity Risk given by the system that on average are too small.

Nevertheless, for SoL verification it is important to verify GNSS integrity performance within a limited time duration. This is especially the case for 'difficult' locations and/or conditions, e.g. near the border of the SBAS' service area or under adverse environmental conditions or system modes, e.g. during a violent ionospheric storm. In order to overcome this problem statistically sound techniques can be applied to GNSS integrity verification, based on the Extreme Value Theory (EVT), which enable extrapolation of the observed distribution's tail into the non-integrity domain [20].

Early investigations [19] based on EVT algorithms developed in Matlab using real SBAS data proved that results from this approach were not only correct, accurate and reproducible, but also very promising due to the limited measurement time duration that was required for obtaining statistically relevant results. Based on GPS/EGNOS data collection over a period of a few months MI probabilities both in the position domain as well as in the signal range domain could be obtained taking into account 2σ confidence levels for MI probabilities Due to the fact that confidence intervals are both related to the minimally needed data collection time period and the capability to prove compliance, optimal determination of confidence levels is important and thus was given special attention. Two approaches of confidence level estimation were studied: using the Gauss-Newton algorithm and the bootstrapping re-sampling approach. Determining confidence levels by both approaches using a few sequences of EGNOS data, it is shown that confidence intervals based on the Gauss-Newton approach are overly optimistic, while the bootstrapping re-sampling approach provides more conservative confidence intervals that is considered adequate.

These promising results positively motivated the Netherlands Space Office to fund the development of a software tool, named GIMAT - GNSS Integrity Monitoring and Analysis Tool, based on these EVT algorithms. GIMAT is developed by a Dutch consortium composed of Delft University of Technology, Integricom, National Aerospace Laboratory NLR and Science & Technology.

GIMAT is now close to its completion. Its first operational application will support the implementation of an EGNOS-based LPV procedure near the Eelde airfield in The Netherlands, testing local GNSS integrity compliance with ICAO SARPs (Standards and Recommended Practices) later this year. Furthermore, plans are currently being made to test GNSS position solution integrity at a location in the North of Scandinavia, near the boundary of the EGNOS service Area, using GIMAT.

The paper presents a short introduction to the mathematical theory of the EVT, the consecutive steps that are made in the software tool to address MI probability estimation and proof of range error distribution overbounding. Full details can be found in [1]. Finally some early results obtained by GIMAT are presented.

1. INTRODUCTION

Of all ICAO GNSS performance requirements the integrity performance requirement is the most stringent one: for instance navigation service levels APV-1, APV-II and LPV-200 require less than 2 x 10^{-7} integrity failure probability per approach [2]. Taking a duration of 150 seconds per approach it turns out that an integrity failure may occur once per 23.8 years (assuming one approach at a time, all the time). Testing a GNSS system by collecting data over such a long period is far from practical and still insufficient from a statistical perspective. So, one needs to apply a method for these integrity tests based on a limited amount of data to be collected within an acceptable observation time. In further discussion we assume the Position Error (PE, or XPE, X being horizontal or vertical) always positive, in line with the Protection Level (PL), while (actual) distributions of XPE/XPL are assessed, also called Safety Index. XPE/XPL > 1 is Misleading Information by definition. If a Gaussian Safety Index (XPE/XPL) distribution could be assumed, then integrity failure statistics could be estimated based on extrapolation of the best fitting normal distribution function. Unfortunately it cannot as the GNSS position

error distribution is not a normal distribution: in general the measured distribution has a heavy tail, see for instance figure 1. As a result a normal distribution would lead to too optimistic integrity estimates. For safety of life applications such as navigation in civil aviation performance of novel navigation aids shall be carefully tested prior to their introduction. The test method presented in this paper and implemented in the developed GIMAT tool is based on EVT. It enables integrity tests providing adequate statistical confidence levels for MI probabilities based on observables collected over a limited period of time.

Given the fact that integrity issues are getting more and more attention due to various reasons, e.g. introduction of GNSS application into safety critical flight operations, the advent of Galileo and the growing ideas on Advanced RAIM introduction [7], while on the other hand currently no tools specifically designed for GNSS integrity verification are available, it was decided to start GIMAT development. The GIMAT tool provides an interesting set of integrity related Key Performance Indicators (KPIs) dedicated to verify GNSS integrity performance versus requirements.

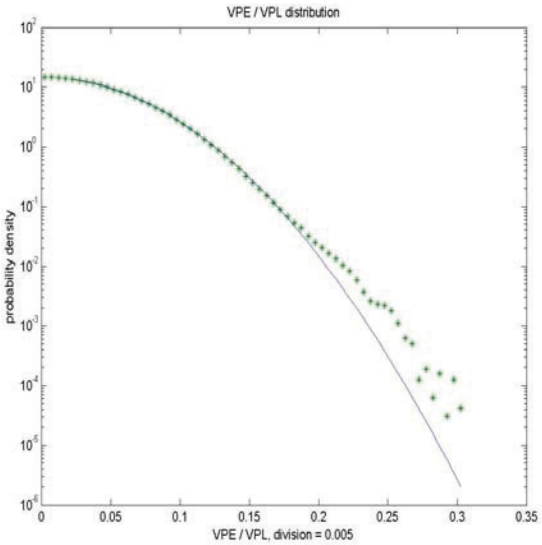

Figure 1: The probability density of the integrity ratio VPE/VPL of a GPS/EGNOS data set, collected during a 3 month period in The Netherlands (6,444,045 epochs). The solid line is a normal distribution, fitting with the actual distribution (*) of VPE/VPL on sample basis.

2. EXTREME VALUE THEORY

Extreme-value theory (EVT) is a recently developed but already well-established and mature field in statistics that provides statistical methods allowing for the estimation of the probability of events that lie beyond the observed range of the data.

The application areas in which EVT has been successfully used are numerous and include hydrology (flood frequency analysis), finance, insurance (insurance and reinsurance risk assessment, claim size assessment, asset price analysis), environmental analysis (ozone level, pollution analysis), meteorology (wind strength and rainfall assessment, earthquake risk assessment) and many engineering areas, e.g. corrosion and fatigue prediction, telecommunication [3].

EVT is applicable regardless of the underlying error distributions of the measurement data, relieving the need for strong a priori assumptions (such as assuming Gaussian error distributions). The properties of the tails of the error distributions can be derived from the measurement data and some weak a priori considerations (which hold for all standard distributions), allowing to meaningfully extrapolate the data into the region of misleading information, even when no (or very limited amounts of) sample values in this region are available. Assuming stationarity of the system, by finding a data-based description of the tails of the error distributions the ICAO conditions can be verified for the tails of the error distributions and the actual protection level can be estimated from the measurement and/or position estimation error data without the need for assuming particular error distributions.

Broadly speaking, there are two principal kinds of models for extreme values. The first group of models, called Generalized Extreme Value theory (GEV) consists of block-maxima models [20]; these are models for the largest observations collected from large samples of identically distributed observations. The GEV cumulative distribution function definition of error e is:

$$H(e)_{K,\sigma,\mu} = \exp\left[-\left(1+K\frac{e-\mu}{\sigma}\right)^{-1/K}\right], \quad for\ 1+K\frac{e-\mu}{\sigma}>0$$

while the probability density is expressed by:

$$f(e)_{K,\sigma,\mu} = \frac{1}{\sigma}\left(1+K\frac{e-\mu}{\sigma}\right)^{-1/K-1}\exp\left[-\left(1+K\frac{e-\mu}{\sigma}\right)^{-1/K}\right], \quad for\ 1+K\frac{e-\mu}{\sigma}>0$$

Here K is called the shape parameter, μ is the location parameter and σ is the scale parameter. For K=0 the distribution is of type I also called Gumbel, for K>0 it is of type II called Fréchet and for K<0 it is of type III also called Weibull. The distribution has a finite left boundary for K>0, a finite right boundary for K<0 and extends to infinity in both directions for K=0. Note that the navigation errors XPE are not bounded toward positive values. As a consequence K is expected not to be negative.

A second group of EVT models contains the peaks-over-threshold (POT) models, which model all large

observations that exceed a high threshold. These POT models are generally considered to be the most useful for practical applications, due to their more efficient use of the (often limited) data of extreme values. Within the POT class of models several styles of analysis exist. Of these the fully parametric models based on the Generalized Pareto Distribution (GPD) have the advantage of being conceptually and implementationally the most simple. The GPD is therefore implemented as the main algorithm within GIMAT in order to model the tails of the error distributions. It is a three-parameter distribution with the following cumulative distribution function of error e:

$$GPD_{\gamma,\beta,\mu}(e) = \begin{cases} 1 & \gamma < 0 \text{ and } e > \mu - \beta/\gamma \\ 1 - \exp\left(-\dfrac{(e-\mu)}{\beta}\right) & \gamma = 0 \\ 1 - \left(1 + \gamma\dfrac{(e-\mu)}{\beta}\right)^{-1/\gamma} & \text{otherwise} \end{cases}$$

In these equations the location parameter μ will be zero when fitting peaks over a threshold. As β (β>0) acts as scaling parameter, the most important parameter is γ, the so-called shape parameter, which largely determines the behaviour of the tail. For positive values of γ, the distribution has a heavy tail, γ=0 corresponds to the exponential distribution, while in the case that γ is negative, the tail has a finite endpoint, that is, the probability density function becomes zero for errors that exceed this endpoint and larger errors cannot occur.

In principle, when the tail of an actual error distribution can be fitted with a GPD cumulative distribution function (cdf) the behavior of the distributions far tail can be estimated and statements on the integrity compliance of GNSS signal in space can be made [1]. There are various issues that need to be addressed though, the most important of which are:

- GNSS system modes
- Data dependencies
- Determination of where the tail starts
- Range domain vs. position domain
- Which KPI to use: overbounding of PE and range errors or (H)MI estimation by cdf extrapolation
- MI vs. HMI and Time To Alert (TTA) aspects
- Confidence levels associated with the statistical KPI determination

Each of these issues will be briefly addressed in the following sections.

3. EVT USED FOR GNSS INTEGRITY TESTS

3.1 Assessment of GNSS System Modes

In general a system mode can be defined as a set of conditions under which the errors in the system can be considered to have a common probability distribution. Different system modes are thus characterised by different error distributions. A system that behaves according to a sequence of different system modes is called non-stationary. Observables from GNSS systems are known to be highly non-stationary and we have to cope with its non-stationarity as it is. This non-stationarity has its largest impact on the data collection. If one collects data during a period of system mode transition this will be reflected in the statistical parameters showing a trend over the measurement period of time, which is obviously not acceptable. In that case one can do one of two things: either one filters one system mode out of the observed data for further processing, e.g. data collected during an ionospheric storm, or one takes the data collection period long enough (longer than the lowest 'system mode frequency wave') in order to average-out all possible system mode effects. This poses certain constraints to the GNSS observables collection and on the outcome of the statistical analysis as one can never be sure that all possible system modes have actually occurred in the data collection period.

Error distributions in the range domain are more sensitive to the existence of different system modes than in position domain, as the position domain errors are the sum of various range domain system mode contributions, which to some extent level out. This means that in range domain one has to be more aware of the fact that the satellites system mode may vary. The different system modes in range domain that are known to have effect on the system's performance will have to be taken into account, e.g. taking into account the fact that the errors depend on the satellite's elevation angle. In principle these system modes need separate analysis for which GIMAT needs data filtering capabilities. However unfortunately several system modes are 'hidden' to the end user, as the system's actual status cannot be known to the user. As a result it will therefore be hard to establish a precise set of equivalent user conditions for which the performance characteristics can be measured in range domain.

For Galileo an additional complicating factor is that the system may show two operational modes at highest system level: Galileo's nominal operation mode and the failure mode, in which at maximum one of the satellites operates in failure.

As the above considerations indicate that knowledge of the exact modes is per definition incomplete (and if known would probably fragment the data too much) the practical approach is to allow mixing, but only as long as there is no knowledge available on basis of which certain user conditions should be regarded as essentially different.

3.2 Data Dependencies

In data analysis where statistical assessment of data samples is involved data dependency plays an important role. For GNSS data dependency would lead to a positive correlation between (simultaneous) errors from different satellites and/or between consecutive samples taken from a single satellite. Concerning correlations between satellites, when these are neglected, this typically leads to an (overly) conservative performance assessment in the position domain [4], due to the fact that common parts of the errors are eventually absorbed into the receiver clock bias.

Traditionally, data dependency in time series is investigated using the autocorrelation function, which for all values k and all pairs of samples is defined as the expected value of the product of two samples at the same time lag (distance) k. Concerning the core of the error distribution, dependency reduction based on autocorrelation function can be used successfully (and will not be further discussed), however for the tail of the distribution another for the tail more suitable approach of dependency reduction is used that avoids unnecessary loss of data.

Correlation between large errors in the tail of the distribution will bring clustering in time of large errors or tail-threshold exceedances. This clustering can be described theoretically by a parameter called the 'extremal index', which is a measure of average cluster size and thus of the dependencies among large errors. Use of the extremal index allows applying extreme value theory to dependent errors without major modifications.

Within GIMAT however another, conceptually simpler, approach is taken that essentially considers clusters of large errors to represent only a single observation while it assumes that observations in different clusters are independent [1]. This approach introduces a risk of eliminating one of two independent events that occur (very) nearby in time, resulting into too optimistic outcomes. Tail events however are (very) rare, so we assume that the chance of nearby independent tail events can be considered negligible.

When focusing on the tail of the error distribution, i.e. selecting those samples for which the error exceeds a defined threshold L_{Tail}, the relatively large error samples should be available at random time instants k_i only. For independent observables the times of their occurrences follow a Poisson distribution. This fact can be used to select a minimum time interval K one needs between two samples to consider them as being independent as follows. Consider all data samples with errors exceeding a certain threshold. When there are N such samples in total, define the fraction of 'large errors' that are at most a time K apart as:

$$V_K = \frac{\text{number of pairs (i,j) for which } \left|k_i - k_j\right| \le K}{\text{N}}$$

In the case of independent samples (Poisson distribution) the expected value $E\{V_K\}$ is linear in K. When one draws a graph of K versus V_K, the graph should therefore approach a straight line above values K_T of K for which independence is obtained, while for smaller K, the graph will be either curving upwards or downwards which indicates some kind of dependence. De-clustering to obtain a set of independent observations (threshold excesses) can thus be performed as follows: use only the largest observations in each cluster, where the large errors are said to belong to the same cluster when they are at most a time K_T apart.

3.3 Determination of Tail Start

In section 3.2 it was identified that the data domain for core and tail needs to be determined, e.g. in order to apply a more optimal dependency reduction, but also in order to obtain optimal GPD parameters β and γ, which distribution fits collected data best asymptotically, i.e. for quantiles approaching infinity. The choice of the threshold L_{Tail} is basically a compromise between:

- Choosing a sufficiently high threshold to make the asymptotic approximation hold,
- Choosing a sufficiently low threshold to obtain sufficient excess data to accurately estimate the parameters γ and β.

Such 'optimal' value of the threshold unfortunately cannot easily be defined objectively: no automatic selection algorithm providing satisfactorily performance is currently available when no a priori assumptions on the data are made. Typically selection is done by human intervention, using a data analysis method as outlined now. After obtaining more experience with this method it possibly may be automated in future.

One of the graphical means to do threshold selection is to depict the 'mean excess plot', (ME) defined as the average of the excesses of the sampled error over different potential thresholds L_{Tail}:

$$ME = E\left\{e - L_{Min} \mid e > L_{Tail}\right\}.$$

Here L_{min} is introduced in order to avoid L=0, which may give problems. The mean excess can be shown to equal [6]:

$$ME = \frac{\beta}{1-\gamma} + \frac{\gamma}{1-\gamma} \cdot L_{Tail} \quad \text{for all } \gamma < 1$$

Note that the ME is thus a linear function of the threshold L_{Tail} as long as the shape parameter remains smaller than one (as the mean excess function is not defined for $\gamma \ge 1$ as the expectation does not exist).

A natural estimate of the Mean Excess ME is given by:

$$ME = \frac{1}{k}\sum_{i=1}^{k}(e_{(i,N)} - L_{Tail})$$

In this equation k is largest subscript i for which $e_{(i,N)} \geq L_{tail}$, $e_{(i,N)}$ is the i-th largest sample (that is, $e_{(1,N)}$ is the largest sample of the error, $e_{(2,N)}$ is the second largest sample, etcetera; $e_{(N,N)}$ is the smallest sample that is observed). For small thresholds, the plot will not be linear as the GPD is not a good approximation and ME will be biased; for large thresholds, the plot will be based on too small a number of data points and ME will be extremely noisy. Between these extremes, there hopefully is a range of thresholds for which the plot is more or less linear and the best choice of a threshold would be at the start (smallest threshold) of this linear area, in order to take as many data points into account as possible. An example is shown in figure 2. Here, L_{tail} would be taken equal to 2. Due to the sparseness of the data for large thresholds, the plot becomes sensitive to noise towards the end of the range, where the theoretically linear curve is compared against a number of small sample realizations. This is typical behaviour, which makes interpretation of the plot often difficult. Yet, the plot is a powerful tool, which is implemented in GIMAT.

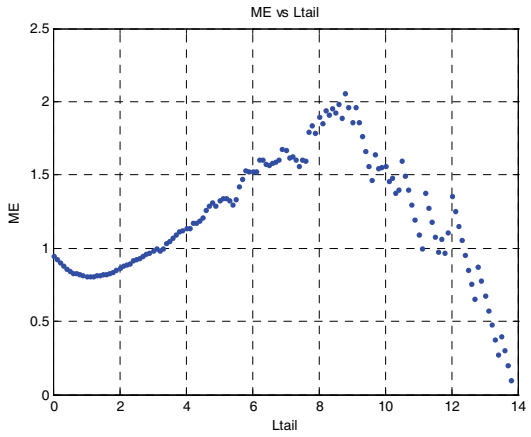

Figure 2: Mean Excess plot for simulated test data set

Because of the difficulties in finding an optimal threshold, it often makes sense to estimate the most important parameter of the GPD, γ for different thresholds L_{Tail}. As soon as the threshold is large enough to make the GPD a good model for the data, the estimated value γ should become more or less stable (although it is of course subject to noise in the data). When the threshold becomes too large, the noise will prevent accurate estimation and the estimated values can show large fluctuations (see figure 3).

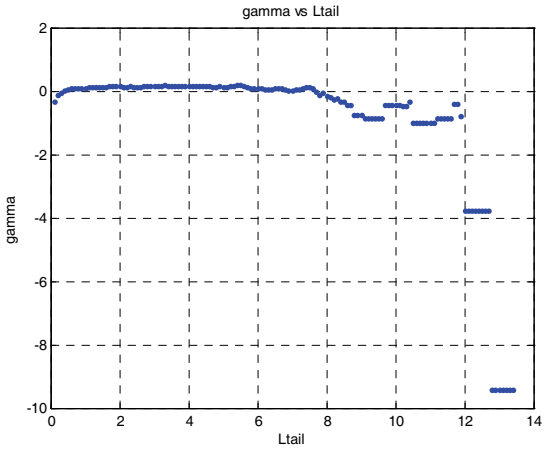

Figure 3: Plot of GPD parameter γ vs L_{Tail} for simulated test data set

One can go one step further and (instead of γ) rather plot the distributional properties of interest, such as the estimated quantiles or exceedance probabilities, against different values of the threshold. This immediately gives a good idea of the sensitivity of the final parameter of interest as a function of the choice of the threshold. Such useful capability is therefore implemented in GIMAT.

3.4 Verification Domain Considerations

In the past it has been debated on which domain to use for integrity verification [8]: the position domain or the range domain? Obviously there is a choice to be made on which verification domain to use for a verification campaign. This choice may depend on the specific purpose but also on the application field of the user one considers. Whatever the selection, both position and range domain integrity verification KPIs are supported by GIMAT. Here, some tradeoffs and suggestions for such a selection are given.

The position domain is the domain in which the user requirements have to be met and thus seems to be a natural choice for system verification. Although the position domain is certainly the most intuitive domain, for which the requirements are defined and best understood intuitively and also checked most straightforwardly, there are some important disadvantages of verification in the position domain.

First, one can verify the system in the position domain either by calculating protection level exceedance probabilities in correspondence with the requirements or, more general for all operations regardless of their requirement, by assessing whether the GNSS error-model distribution overbounds the actual error distribution. In both cases, the outcomes of the procedure will in general depend on the particular mix of geometries for which the verification is performed, as it is possible that system

integrity is within the requirement for certain geometries, while for others it is not.

Secondly, using the position domain instead of the range domain involves loss of information on the mechanism of the providence of integrity, which is in fact working on the range level. As a result of this, it will be harder to distinguish between different system modes that each individual satellite is in. While the position error is a constantly changing linear combination of range errors (RE) (due to geometry changes) it is practically not feasible to consider each geometry as a separate system mode. When the position domain is used for verification purposes, one should limit the distinguished system modes to only those that affect all satellites simultaneously, such as for example ionospheric storms. Resuming, system mode specific integrity checks can only be done well in range domain.

Summarizing, the advantage of verifying the system in the position domain is that:

- the position domain allows for a relatively straightforward check (no precise range determination required) on whether the requirements are met and is less conservative than the range domain.

Disadvantages include:

- the verification outcome will depend on the particular mix of geometries that have been used in obtaining the position errors;
- system modes that are significantly different in their range-domain performance will be much harder to distinguish in the position domain;
- for Galileo, the position domain involves a multi-dimensional error vector rather than a scalar error to be analyzed.

While it is realized that the requirements are essentially stated in the position domain, range domain analysis is included in order to reveal optimally what is going on in the system. It can also be stated that range domain overbounding automatically implies the presence of position domain overbounding for all possible satellite geometries as long as the actual (but unknown) range error distributions are symmetric, unimodal and only slightly biased [9].

The advantages of verification in the range domain can be summarized as follows:

- range domain integrity determination (overbounding and integrity risk) does not depend on satellite geometry; the results are therefore valid for all possible geometries and not just for the particular mix of geometries that have been observed during the verification campaign;

There are disadvantages for the range domain as well:

- range domain overbounding requires implicit assumptions on the range error distribution (symmetry, unimodality, nearly unbiasedness) that are quite difficult to check;
- range domain integrity verification is (overly) conservative with respect to the position domain counterpart;
- in the position domain the reference position is generally known with higher accuracy than the reference range in range domain.

When working in the range domain, prior to using observables for KPI determination, the (e.g. SBAS-)-corrected range ρ_{corr} and its reference, the true geometric range ρ_{true} per GNSS satellite, shall be determined. The calculation of ρ_{corr} is done in GIMAT in line with [10] using the SBAS message set as input, while ρ_{true} is determined using IGS final orbits including precise ephemeris (position X,Y,Z) of GPS satellites for corresponding timeline using IGS (sp3) data files. An interpolation method (see e.g. [11]) is necessary to convert the 15 minute ephemeris stamps in sp3 file format into 1 Hz samples for the used GPS satellites.

3.5 Distribution Overbounding vs. Misleading Information Risk Extrapolation

In line with the statements on position domain vs. range domain, GIMAT supports both integrity verification based on the overbounding concept [1] as well as on MI probability estimation, i.e. based on cumulative distribution function (cdf) extrapolation into the non-integrity domains. Overbounding is a concept that describes the relation between the actual error distribution and a selected model distribution, generally a Gaussian distribution. The Protection Level relies on this overbounding distribution and should be computed such that it overbounds the actual error distribution [8]. Preference for either one of those two KPIs may be based on the specific purpose of study but also on the background of the user: while a system engineer responsible for the correct operation of the GNSS system may be more interested in overbounding (because of its capability to verify protection levels integrity irrespective of the actual integrity requirements specifications), the end user will be more interested in GNSS compliance with requirements for his specific application including associated confidence levels. For this reason both KPIs, overbounding of PE and RE and MI probability estimation, are implemented in GIMAT.

The approach of overbounding-based integrity verification is the following. First some preparatory steps shall be made:

- In case of range domain the ρ_{corr} and ρ_{true} shall be determined and normalized

- The core- and tail-part of the data distribution shall be identified
- Both core- and tail-data shall be down sampled in order to remove data dependencies

The resulting datasets will subsequently be tested on overbounding capability (which approach may be different for core and tail). For this testing of 'first order stochastic dominance' (overbounding) it shall either be proved that for both the core and the tail of the distribution the model distribution does dominate the actual distribution or that the model distribution does not dominate the actual distribution. In literature several of those stochastic dominance tests are worked out, e.g. [12, 13, 14, 15].

For testing overbounding of the distribution's core within GIMAT the Kaur, Rao and Singh (KRS) test is adopted [1, 12]:

$$T_N(L) = \frac{\sqrt{N}(2\bar{\Phi}(L) - \hat{\bar{F}}_{actual}(L))}{\sqrt{\hat{\bar{F}}_{actual}(L)(1-\hat{\bar{F}}_{actual}(L))}}$$

$$T_{KRS} = \min_{L_{Min} \le L \le L_{Tail}} \left(T_N(L)\right)$$

$$T_{KRS} < z_\alpha \Rightarrow no\ overbounding$$
$$T_{KRS} \ge z_\alpha \Rightarrow overbounding$$

In which:
N is the number of down sampled core samples
L is the downsampled normalized error value data (sorted)
$\bar{\Phi}(L)$ (or $1-\Phi(L)$) is the complement of the cumulative standard normal (or Gaussian) distribution function.
$\hat{\bar{F}}_{actual}(L)$ is the complement of the empirical cumulative distribution function of L:

$$\hat{\bar{F}}_{actual}(L) = \frac{N \le L}{N}$$

L smaller than L_{MIN} should be ignored (as L=0 may give problems). In the test, the threshold is given by z_α, which is the $(1-\alpha)$ quantile of the standard normal distribution.

When testing the tail on overbounding using KRS statistics the bootstrapping resampling approach [16, 17] is best used in addition. Bootstrapping is based on re-sampling the data set in order to generate multiple datasets that can be used to reveal information on the statistical properties of quantities derived from the original dataset. A great advantage of bootstrapping is its simplicity, combined with the lack of strong assumptions needed on the distributions involved. However simplicity is achieved at the expense of a (much) larger computational effort. In order to determine P_{Over_tail} the following equation is applicable:

$$P_{Over\_tail} = 1 - \left[\frac{\sum_{1}^{n\_boot} bool\left(if \left\langle \min(2\bar{\Phi}(L_{check}) - \hat{\bar{F}}_{actual}(L_{check})) < 0 \right\rangle true \right)}{n\_boot} \right]$$

In which:

$$\hat{\bar{F}}_{actual}(L_{check}) = \frac{N_{L_{Tail}}}{N} \overline{GPD}_{\hat{\gamma},\hat{\beta}(L)}\left(L_{check} - L_{Tail}\right)$$

$\hat{\bar{F}}_{actual}$ is the probability that the error exceeds the chosen tail threshold L_{Tail}.
L is tail data set
$N_{L_{Tail}}$ is the number of L.
N the total number of samples (core + tail).
$\overline{GPD}$ is the complement of the cumulative GPD distribution function for each new tail data set
L_{check} is a vector of extrapolated tail data used to check whether tail data is overbounded in regions for Φ of e.g. 1-1e-8 (or 1e-8 for $\bar{\Phi}$).
$bool\left(if \left\langle \min(2\bar{\Phi}(L_{check}) - \hat{\bar{F}}_{actual}(L_{check})) < 0 \right\rangle true \right)$ is a non-overbounding criteria resulting in "1" if true and "0" for false.
The final result P_{Over_tail} should be larger than $(1-\alpha)$ confidence level for tail-overbounding to be proved.

Integrity risk determination by extrapolation of the obtained GPD function can be calculated quite straightforward. First L_{Tail} must be determined, where L=HPE/HPL for the horizontal situation and L=VPE/VPL for the vertical case. Following, when the corresponding GPD parameters $\hat{\gamma}$ and $\hat{\beta}$ are determined. Finally the probability of Misleading Information can be calculated for L=1 using

$$P_{MI} = \overline{GPD}_{\gamma,\beta}(L) = 1 - GPD_{\gamma,\beta}(L) = \left(1 + \frac{\hat{\gamma}}{\hat{\beta}} L\right)^{-1/\hat{\gamma}}$$

Applicable confidence levels of the exceedance probability can be obtained again by the bootstrapping approach. In this way a distribution of exceedance probabilities is obtained from which e.g. the P_{MI95} (exceedance probability including 95% confidence level) can be derived.

3.6 MI vs HMI and Time To Alert (TTA) aspects

In the previously described approach for determining non-integrity probability, clearly MI probability is determined here, where e.g. ICAO requirements should be interpreted as applicable for hazardously misleading information (HMI) probability, because MI events with XPL > XAL are already flagged as non-available and MI events where (XPE > XPL) ∩ (XPE < XAL) are not leading to a safety incident. In addition, an HMI event is only recognized as such when no alert was generated within TTA seconds. So, hazardously misleading information events occur when during an operation the following conditions are true (for either the horizontal or vertical case) simultaneously:

HMI: ((HPE > HAL) ∩ (HPL < HAL)) ∩ no timely alert

and similar for the vertical case, where MI occurs when *MI: HPE > HPL* and vertical case similar.

Question is therefore how to interpret P_{MI} results where P_{HMI} is requested? The following remarks can be made here.

First, since

$$(HPE > HAL) \cap (HPL < HAL) \rightarrow HPE > HPL$$

it follows that the HMI events are a subset of MI events. This is also shown clearly by the Stanford diagram (see figure 4): only the lower right square area is considered as HMI.

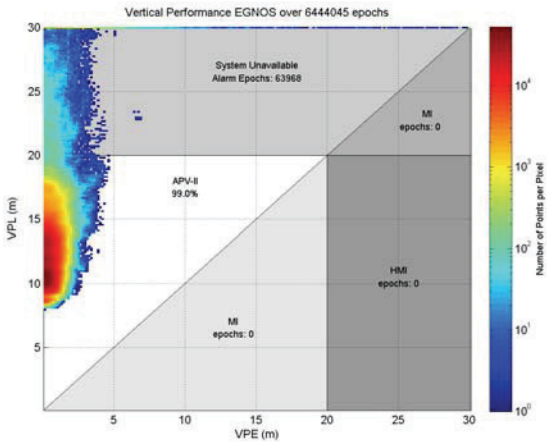

Figure 4: Stanford diagram for the aeronautical service level APV2 (Vertical Alarm Limit = 20m), populated with 3-months EGNOS observables collected in The Netherlands

As a result P_{MI} is an (overly) conservative probability estimate with respect to the requested P_{HMI}. Therefore, in case MI compliance to requirements can be proved, then the dataset will satisfy the HMI requirement as well.

As a second resort, in the event compliance cannot be proved for MI, the test can be made less conservative by filtering the input dataset such that samples that could lead to MI but not to HMI will be removed as much as possible. In practice this means all collected epochs that cannot contribute to HMI, i.e. all samples for which XPL > XAL (for leading to unavailability) are removed from the dataset prior to integrity analysis.

Third, in addition filtering input data based on Alerts and the associated Time To Alerts can be taken into account. The SBAS and Galileo ground-segments need some time to detect large ranging errors and transmit this information to the user. It is decided to deal with this by disregarding the samples over the TTA time period (generally 6 seconds) before the integrity flag is raised by the ground segment to ensure that these samples don't penalize the integrity analysis.

These measures enable the tool not to be overly conservative and produce results that are more in line with HMI probability estimation as requested.

3.7 Confidence levels associated with statistical KPI determination

When doing very precise KPI determination, such as is done in this case, determining confidence intervals and confidence levels associated with the determination are inevitable. Usually, the confidence interval of interest is symmetrically placed around the mean, so a 50% confidence interval for a symmetric probability function would be the interval [−a,a] such that:

$$\frac{1}{2} = \int_{-a}^{a} P(x)dx$$

The confidence interval is not a system parameter, however generally relates with quality and quantity of the measurement data set. Various methods for estimating confidence intervals exist, the applicability of which depends on the kind of application for which it is used. Most straightforward would be to repeat the experiment many times and determine the confidence interval based on the distribution of outcomes. In general though, this approach is not very practical or even impossible. Prior to adoption a method for confidence interval determination its performance for the specific application should be tested. For confidence levels associated with integrity KPI determination we have investigated two possible methods: the Gauss-Newton algorithm and the re-sampling bootstrapping approach. The Gauss-Newton iteration is a natural choice: for curve fitting of the GPD and GEV functions, determining its shape parameters (γ, β, μ and K, σ, μ respectively) based on the set of measured tail samples the Gauss-Newton 'maximum likelihood method' is used. This iteration produces confidence intervals and levels for the shape parameters as a by-product [18]. The alternative method was bootstrapping, its advantage being its great simplicity together with lack of need for strong assumptions: it is straightforward to

derive estimates of standard errors and confidence intervals for complex estimators of properties of the distribution. Moreover, it is an appropriate way to control and check the stability of the results. As test input a 3-month period of collected GPS/EGNOS observables was cut into 5 consecutive time periods. For each of the data sets the shape parameters were determined together with their 95% confidence intervals. These confidence intervals were calculated both using Gauss-Newton and bootstrapping (using 1000 re-sampled data sets). The results are shown in figure 5. It is expected that for each of the three GEV parameters there exist a value that satisfies the 95% confidence intervals for all five data subsets subsequently. From figure 5 it is apparent that this is not the case for the Gauss-Newton iteration results. Therefore the Gauss-Newton confidence interval determination seems to be too optimistic and as a result the bootstrapping approach was decided to be implemented into GIMAT.

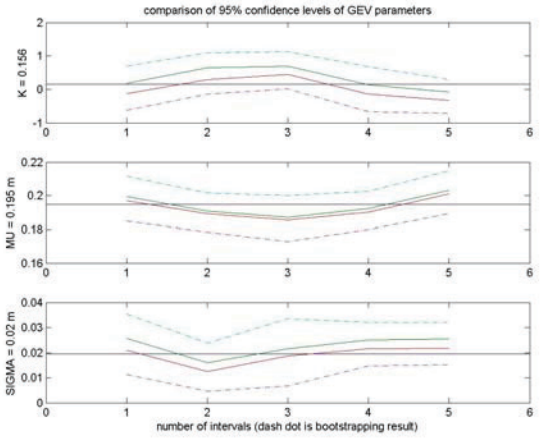

Figure 5: Comparison of bootstrapping versus Gauss-Newton algorithm for confidence intervals of GEV shape parameters as function of five data subsets

4. EARLY RESULTS

All developed algorithms are currently consolidated and integrated into the GIMAT tool, providing a graphical user interface and enabling management of data sets and inspection of intermediate and end results by means of various figures. Some interesting preliminary results (although it concerns GEV results instead of not yet available results from the more promising GPD method) obtained during algorithm development can be presented already [19].

Using the GEV/block-maxima approach the position domain integrity performance of the GPS/EGNOS combination was tested in the Netherlands, which is about at the heart of the EGNOS coverage area. For determination of 95% confidence intervals for the Safety Index PE/PL the bootstrapping approach was used. Figure 6 shows the vertical probability density function in the position domain based on 3 months collected observables. When integrating the area below the pdf and starting beyond x=1 (VPE/VPL=1), the $P_{MI,V}$ is obtained in the vertical direction. When including the upper part of the confidence interval in figure 6 as well the $P_{95,MI,V}$ is obtained. This $P_{95,MI,V}$ was found to be $8 \cdot 10^{-8}$ satisfying ICAO CAT 1 requirements. Thus indications were found that based on this 3-month data set the GPS/EGNOS integrity performance, including the upper part of the confidence interval satisfies ICAO requirements for the measured location in the Netherlands, assuming that all applicable system modes are balanced represented in the data set (see Table 1 below). As can be seen, of the two components horizontal and vertical, the vertical one is the most critical.

ICAO Aeronautical Service level	$P_{MI,H}$ (95% conf level)	$P_{MI,V}$ (95% conf level)	$P_{HMI,H}$ (95% conf level)	$P_{HMI,V}$ (95% conf level)	Req (H+V comb)
APV-I (HAL=40 m, VAL= 50m)	$6 \cdot 10^{-10}$	$8 \cdot 10^{-8}$		$2 \cdot 10^{-9}$	$2 \cdot 10^{-7}$
APV-II (HAL= 40m, VAL= 20m)	$6 \cdot 10^{-10}$	$8 \cdot 10^{-8}$	$8 \cdot 10^{-11}$	$1 \cdot 10^{-8}$	$2 \cdot 10^{-7}$
CAT-I (HAL= 40m, VAL= 15m)	$6 \cdot 10^{-10}$	$8 \cdot 10^{-8}$	$8 \cdot 10^{-11}$	$6 \cdot 10^{-8}$	$2 \cdot 10^{-7}$

Table 1: GPS/EGNOS Integrity performances per approach, the Netherlands, based on 3-month collected data set (GEV/bootstrapping)

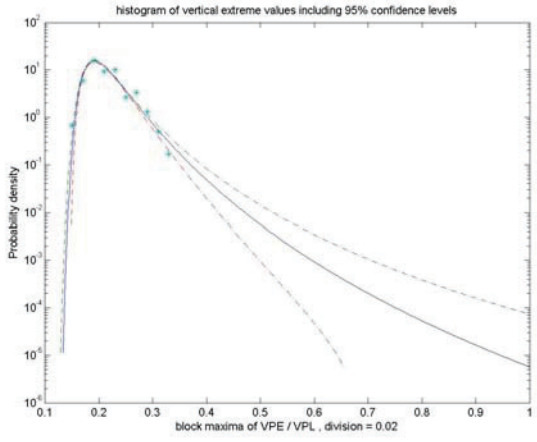

Figure 6: Histogram of VPE/VPL daily block maxima together with the fitting GEV probability density function including confidence intervals (dashed lines) based on 3-month GPS/EGNOS observables.

5. CONCLUSIONS

With GIMAT a tool is now available for statistically sound integrity performance verification using receiver output data to be gathered during a test campaign of

limited duration. The method is based on the Extreme Value Theory allowing a reliable statement of overbounding of the position error and range error actual distribution by a probability density function and extrapolation of the fitted GPD cumulative probability function toward MI and HMI regions. The tool is interesting both from the end-user perspective and the GNSS system builder and operator point of view. In addition for SoL certification decision making now it can be determined whether the most important positioning Key Performance Indicator (KPI), namely integrity, for a given GNSS system solution, whether it will be GPS+SBAS, GPS+RAIM, GBAS, Galileo or Advanced RAIM, satisfies given requirements for a specific navigation service level.

ACKNOWLEDGMENTS

The authors would like to thank the Netherlands Space Office. Without their funding this tool development would not have been possible.

REFERENCES

[1] Galileo and EGNOS Integrity Validation, Ober P.B., NSO-GIMAT-INT-TN-R-001, (GIMAT Internal Doc.), 2011

[2] ICAO, ANNEX 10, Aeronautical Telecommunications, Volume 1 (Radio Navigation Aids), Amendment 84, Published July 20, 2009, effective November 19, 2009.

[3] Statistical Analysis of Extreme Values: with Applications to Insurance, Finance, Hydrology and other Fields, R.D. Reiss, M. Thomas; Birkhauser, 2001.

[4] Validation of the WAAS MOPS Integrity Equation, Walter T., Hansen A., Enge P., Proceedings of the 55th Annual Meeting of The Institute of Navigation, 2002, pp766 – 772.

[5] On Nonparametric Measures of Dependence for Random Variables, Schweizer B., Wolff E.F., Annals of Statistics. 1981:9, pp879 – 885.

[6] Lecture notes: Extreme quantile and tail estimation, de Haan L., 2003

[7] A Proposal for Multi-Constellation Advanced RAIM for Vertical Guidance, Juan Blanch et al., Proceedings of ION GNSS 2011, September 2011, Portland, Or, USA.

[8] Statistical Validation of SBAS Integrity, Ober P.B., de Haan L., Li D., Proceedings of the European Navigation Conference GNSS 2004, Rotterdam, The Netherlands.

[9] An application of Gaussian overbounding for the WAAS Fault Free Error Analysis, Schempp T.R., Rubin A.L., Proceedings of ION-GPS 2002.

[10] WAAS Minimum Operational Performance Specifications (MOPS), RTCA document DO-229D

[11] InsideGNSS GPS EASY Suite II; 16-Comparing GPS Precise and broadcast Ephemerides, K. Borre, May 2010.

[12] Testing for Second-Order Stochastic Dominance of Two Distributions, Kaur A., Prakasa Rao B.L.S., Singh H., Econometric Theory, 1994 December; 10(05):849-866

[13] Tse YK, Zhang X., A Monte Carlo investigation of some tests for stochastic dominance, Journal of Statistical Computation and Simulation, 2004;74(5):361-378

[14] Modification of the Kolmogorov-Smirnov Statistic for Use with Correlated Data, Weiss M.S., Journal of the American Statistical Association, 1978;73(364): 872-875

[15] Consistent Testing for Stochastic Dominance: a subsampling approach, Linton O., Maasoumi E., Wang Y.J., Centre for Microdata Methods and Practise, Institute for Fiscal Studies; 2002, CWP03/02.

[16] Resampling methods for Dependent Data, Lahiri S.N., Springer Series in Statistics, Springer, 2003.

[17] Bootstrap Methods and their Application, Davidson A.C., Hinkley O., Cambridge University Press, 1997

[18] Practical Methods of Optimization, Fletcher R., Springer ISBN 0387987932, 1987.

[19] The Generalized Extreme Value Statistical Method to Determine the GNSS Integrity Performance, Kannemans H., NLR-TP-2010-491, October 2010.

[20] Extreme Value Distributions: Theory and Applications, Kotz, S. and Nadarajah, S., Imperial College Press, 2000