The Generalized Method of Wavelet Moments with eXogenous inputs

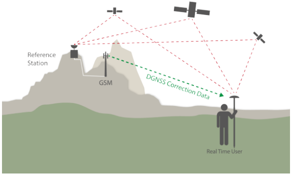
Lionel Voirol [1]

Joint work with:

Davide A. Cucci [1]
Gaël Kermarrec [2]
Jean-Philippe Montillet [3, 4]
Stéphane Guerrier [1, 5]

[1] Geneva School of Economics and Management, University of Geneva, Switzerland
[2] Institute for Meteorology and Climatology, Leibniz University Hannover, Germany
[3] Institute Dom Luiz (IDL), University of Beira Interior, Portugal
[4] Physikalisch-Meteorologisches Observatorium Davos/World Radiation Center (PMOD/WRC), Davos, Switzerland
[5] Faculty of Science, University of Geneva, Switzerland

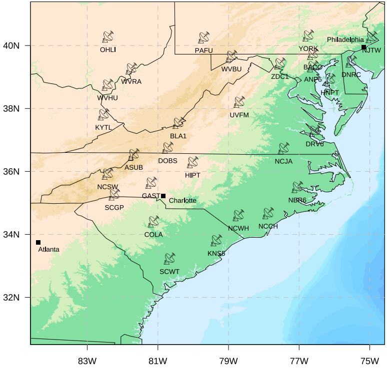
Research context and objectives

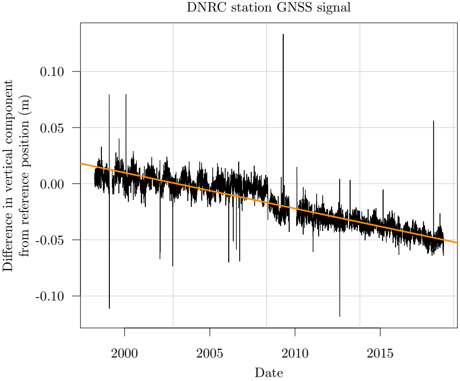


- **The Global Navigation Satellite System (GNSS)** is an important tool to observe and model **geodynamic processes** such as post-glacial rebound, hydrological loading or crustal deformations. GNSS signals also have **important practical geopositioning** applications such as for example when used with differential methods to improve GNSS position accuracy.
- Due to the considerable computational resources required, GNSS time series analysis is commonly performed with daily observations. Yet, in many cases, it is preferable to analyze data hourly or even on a minute-by-minute basis.

Research Objective: Analysis of large network of GNSS stations is extremely computationally intensive and **our objective is to develop an alternative estimator that is considerably more computationally efficient with a reasonable loss in efficiency**.

Example of a GNSS network and of a GNSS time series signal





- In many applications, the primary objective is to obtain a **precise estimation of a specific parameter**, with a particular emphasis on the **signal's trend**, and to accurately assess the associated **uncertainty of this parameter**.

The model

- A general formulation of the models typically used to model GNSS signals can be expressed as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ denotes the response variable of interest (i.e. the vector of GNSS observations), $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is a fixed design matrix, $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is a vector of unknown constants and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a vector (mean zero) of residuals.

- We assume that $\boldsymbol{\varepsilon}_t$ is a **strictly stationary process** with $\boldsymbol{\varepsilon} \sim \mathcal{N}\left\{\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\gamma}_0)\right\}$, where $\boldsymbol{\Sigma}(\boldsymbol{\gamma}_0) > 0$ and that **it depends on the unknown parameter vector $\boldsymbol{\gamma}_0 \in \mathbb{R}^q$**. This matrix **does not have** a **block diagonal structure** neither a **Toeplitz structure**.
- The noise structure is generally modelled with **latent composite stochastic models** which often consider **long-memory stochastic processes.**
- Hence, we define

$$\boldsymbol{\theta}_0 := \begin{bmatrix} \boldsymbol{\beta}_0^{\mathrm{T}} & \boldsymbol{\gamma}_0^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} \in \boldsymbol{\Theta} \in \mathbb{R}^{p+q}$$

as the vector of parameters of the model.

Maximum Likelihood Estimator (MLE)

- The likelihood function for a generic $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is simply given by

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}} \boldsymbol{\Sigma}(\boldsymbol{\gamma})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right\}\left[(2\pi)^n \det\{\boldsymbol{\Sigma}(\boldsymbol{\gamma})\}\right]^{-1/2},$$

  allowing to define the Maximum Likelihood Estimator (MLE) for $\boldsymbol{\theta}_0$ as

$$\hat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{\boldsymbol{\beta}}^{\mathrm{T}} & \widehat{\boldsymbol{\gamma}}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmax}} \quad L(\boldsymbol{\theta} \mid \boldsymbol{Y}) \tag{1}$$

- Using standard regularity conditions, the MLE is **asymptotically normal** and **asymptotically efficient**.
- Solving (1) require to evaluate the likelihood function a large number of time where each evaluation involves the inversion of the $n \times n$ matrix $\boldsymbol{\Sigma}(\boldsymbol{\gamma}_0)$. This operation has **a computational complexity of order $\mathcal{O}(n^\delta)$ where $\delta \in [2, 3]$** depending on the considered algorithm.
- In practice, the analysis of complex geodynamic processes requires to estimate signals from hundreds to thousands of GNSS stations (He et al., 2021) which record daily observations over decades and where different noise models must be tested.
- This procedure which has to be performed routinely becomes **impractical due to the large amount (e.g., weeks) of processing time required** (He et al., 2019; Bos et al., 2020).

- We propose to use of a new **two-step statistical procedure**, which considers a **Generalized Least Squares (GLS)** approach combined with the **Generalized Method of Wavelet Moments (GMWM)** proposed in Guerrier et al., 2013.
- The proposed estimator is an iterative method and the number of iteration j can be used to balance the statistical properties and computational cost.
- We denotes this estimator as the **GMWMX** in reference to the **Autoregressive–moving-average model with eXogenous inputs model (ARMAX)**.
- We first define the GMWMX estimator $\tilde{\boldsymbol{\beta}}$ of the parameters $\boldsymbol{\beta}_0$ which corresponds to the Generalized Least Square estimator.

$$\tilde{\boldsymbol{\beta}}(\boldsymbol{\Sigma}) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \{\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \{\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\} = \left(\boldsymbol{X}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{y}$$

- We define $\boldsymbol{\varepsilon}(\boldsymbol{\beta}) = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ and its natural estimator based on $\tilde{\boldsymbol{\beta}}$, $\tilde{\boldsymbol{\varepsilon}}_i = \boldsymbol{\varepsilon}_i(\tilde{\boldsymbol{\beta}}) = \boldsymbol{y}_i - \boldsymbol{X}_i^{\mathrm{T}} \tilde{\boldsymbol{\beta}}$

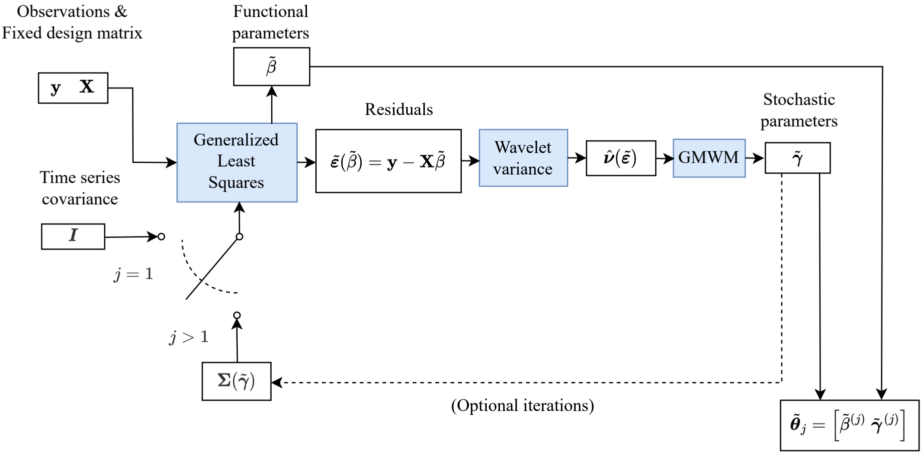
The GMWMX

- We then consider a **GMWM methodology** in order to construct a **computationally efficient estimator of γ_0** using the Wavelet Variance (WV) of the residuals $\boldsymbol{\varepsilon}(\boldsymbol{\beta})$.
- The GMWM (Guerrier et al., 2013) is a computationally efficient moment-based estimator which exploits the mapping between the theoretical Wavelet Variance (WV) implied by a model and the empirical WV estimated on a signal.
- We define $\boldsymbol{\nu}(\boldsymbol{\gamma})$, the WV implied by the estimated model and $\widehat{\boldsymbol{\nu}}\,(\boldsymbol{\beta})$ which corresponds to the estimated Haar WV computed on $\boldsymbol{\varepsilon}\,(\boldsymbol{\beta})$, in order to estimate the vector of parameters of interest $\boldsymbol{\gamma}_0$.
- We define

$$\tilde{\boldsymbol{\gamma}}(\boldsymbol{\beta}) = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \{\hat{\boldsymbol{\nu}}(\boldsymbol{\beta}) - \boldsymbol{\nu}(\boldsymbol{\gamma})\}^{\mathrm{T}} \boldsymbol{\Omega} \{\widehat{\boldsymbol{\nu}}(\boldsymbol{\beta}) - \boldsymbol{\nu}(\boldsymbol{\gamma})\},$$

  where $\boldsymbol{\Omega}$ is an appropriate (possibly estimated) positive-definite weighting matrix.
- The **computational bottleneck** of this procedure corresponds to the computation of the empirical WV which has **a computational complexity of order $\mathcal{O}(n\log(n))$.**

GMWMX: Iterative algorithm (Flowchart)



GMWMX: Iterative algorithm

- We define the estimator resulting from j iterations and hence using an updated estimator of $\boldsymbol{\Sigma}(\boldsymbol{\gamma}_0)$ as $\tilde{\boldsymbol{\theta}}^{(j)}$.
- Starting at $j = 1$ with $\boldsymbol{\Sigma}^{(0)} = \boldsymbol{I}$, we define

$$
\begin{aligned}
\widetilde{\boldsymbol{\beta}}^{(j)} &= \left[\boldsymbol{X}^{\mathrm{T}} \left(\boldsymbol{\Sigma}^{(j-1)}\right)^{-1} \boldsymbol{X}\right]^{-1} \boldsymbol{X}^{\mathrm{T}} \left(\boldsymbol{\Sigma}^{(j-1)}\right)^{-1} \boldsymbol{y}, \\
\widetilde{\boldsymbol{\gamma}}^{(j)} &= \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \left\{\hat{\boldsymbol{\nu}}\left(\widetilde{\boldsymbol{\beta}}^{(j)}\right) - \boldsymbol{\nu}(\boldsymbol{\gamma})\right\}^{\mathrm{T}} \boldsymbol{\Omega} \left\{\hat{\boldsymbol{\nu}}\left(\widetilde{\boldsymbol{\beta}}^{(j)}\right) - \boldsymbol{\nu}(\boldsymbol{\gamma})\right\}, \\
\boldsymbol{\Sigma}^{(j)} &= \boldsymbol{\Sigma}\left(\widetilde{\boldsymbol{\gamma}}^{(j)}\right) = \operatorname{var}\left(\boldsymbol{y} \mid \widetilde{\boldsymbol{\gamma}}^{(j)}\right).
\end{aligned}
\tag{2}
$$

- The resulting estimator is hence denoted by: $\widetilde{\boldsymbol{\theta}}^{(j)} = \left[\begin{array}{cc} \widetilde{\boldsymbol{\beta}}^{(j)\mathrm{T}} & \widetilde{\boldsymbol{\gamma}}^{(j)\mathrm{T}} \end{array}\right]^{\mathrm{T}}$
- We denote the estimator defined in Eq. (2) with one or two iterations as the GMWMX-1 and the GMWMX-2, respectively.
- Under arguably weak conditions (Guerrier et al., 2013; Guerrier et al., 2022), the resulting estimator is **consistent** and **asymptotically normal**.
- Moreover, it can be shown that $\widetilde{\boldsymbol{\beta}}^{(j)}$ is asymptotically optimal for all $j \geqslant 2$ in the sense that

$$
\lim_{n \to \infty} \operatorname{var}\left\{\sqrt{a_n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)\right\} - \operatorname{var}\left\{\sqrt{a_n}\left(\tilde{\boldsymbol{\beta}}^{(j)} - \boldsymbol{\beta}_0\right)\right\} = 0
$$

where $\{a_n\}_{n \in \mathbb{N}}$ is a diverging sequence of positive numbers such that $\sqrt{a_n}$ corresponds to the asymptotic rate of convergence of $\widehat{\boldsymbol{\beta}}$.

Simulation Studies: Description & Settings

- We evaluate the performance of the GMWMX-1 and GMWMX-2 estimators with respect to the MLE implemented in the open source software `Hector v1.9` (Bos et al., 2008) which represents the fastest available implementation of the MLE for these type of models.
- We generate signal of different lengths of GNSS daily position time series, i.e., 2.5, 5, 10, 20 and 40 years and consider 10 % of missing observations for each simulated signal which corresponds approximately to the estimated median number of missing data of publicly available datasets (Bos et al., 2013).
- We fix the parameters of the model by considering values which are representative of the estimated parameters on real GNSS time series signal.
- All our simulations are based on $B = 10^3$ Monte-Carlo replications.

Model considered

- A common formulation of of the model is given by He et al., 2017, which can we expressed as follows for the i-th component of the vector $\boldsymbol{X}\boldsymbol{\beta}_0$:

$$\boldsymbol{X}_i^T\boldsymbol{\beta}_0 = a + b\left(t_i - t_0\right) + \sum_{j=1}^{2}\left[c_j \sin\left(2\pi f_j t_i\right) + d_j \cos\left(2\pi f_j t_i\right)\right] + \sum_{k=1}^{n_g} g_k H\left(t_i - t_k\right),$$

  where:
  - a is the initial position at the reference epoch t_0
  - b is the trend parameter
  - c_j and d_j are the periodic motion parameters (where f_j is the frequency of the sinusoidal and $j = 1$ and $j = 2$ represent the annual and semi-annual seasonal terms, respectively).
  - The offsets term models earthquakes, equipment changes or human intervention, in which g_k is the magnitude of the change at epochs t_k, n_g is the total number of offsets, and $H(x)$ is the Heaviside step function $H(x) := \begin{cases} 1, & x > 0 \\ 0, & x \leq 0 \end{cases}$
- Regarding the stochastic model considered for $\boldsymbol{\varepsilon}$, we consider **the sum of a White noise and a Matérn process**, where the autocovariance function of the Matérn process with parameter α, λ and σ^2 is given by:

$$a(h) = \frac{2\sigma^2}{\Gamma(\alpha - 1/2)2^{\alpha-1/2}}|\lambda h|^{\alpha-1/2}\mathcal{K}_{\alpha-1/2}(\lambda|h|)$$

where $\mathcal{K}_\omega(x)$ is the modified Bessel function of the second kind of order ω.

Simulation Studies: Computational gain

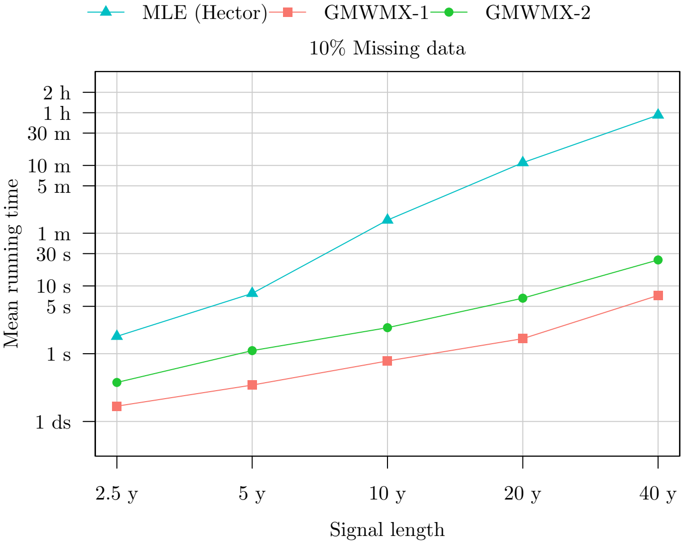


Figure: Mean running time of the MLE, the GMWMX-1 and the GMWMX-2 as a function of the sample size.

Simulation Studies: Point estimation

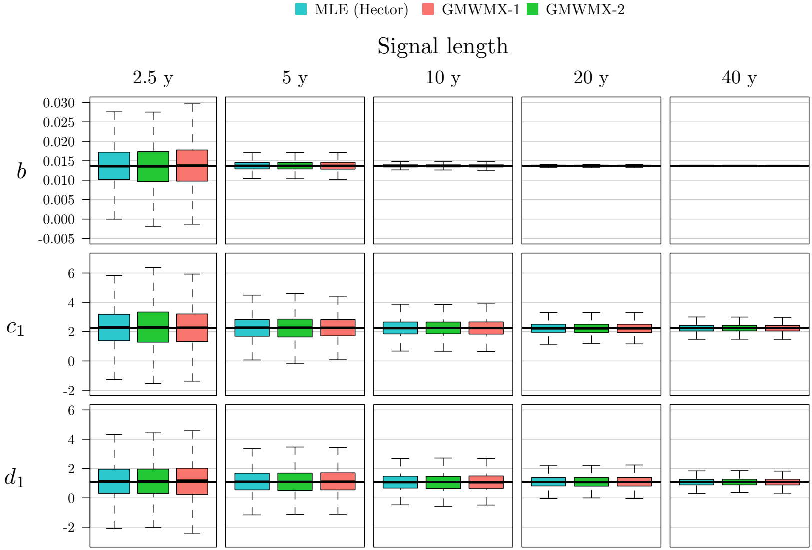


Figure: Boxplots of the estimated parameters of the model with the GMWMX-1, the GMWMX-2 and the MLE

Simulation Studies: Point estimation

Figure: Boxplots of the estimated parameters of the model with the GMWMX-1, the GMWMX-2 and the MLE

Simulation Studies: Inference

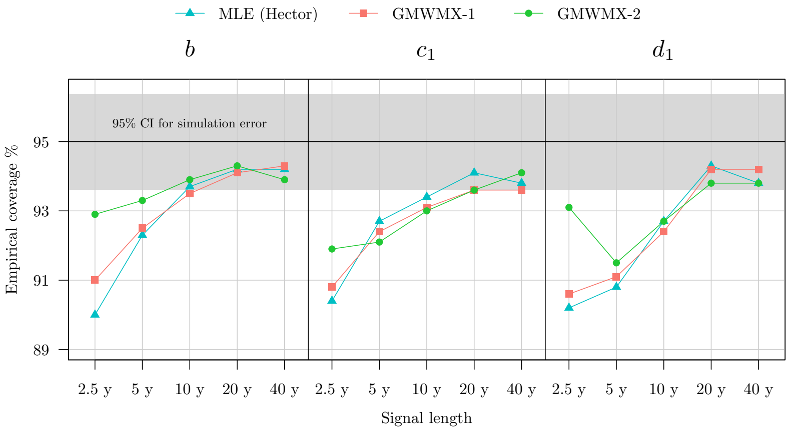


Figure: Empirical coverage of the confidence intervals at level $1 - \alpha = 0.95$ for the parameters b, c_1 and d_1 for GMWMX-1, GMWMX-2 and the MLE as a function of sample size. The grey area represents a 95% confidence interval of the simulation error.

Case Study

- We apply our method to daily GNSS coordinate time series. We use measurements from **a small network of 33 continuously operating GNSS receivers** distributed over the east coast of the USA.
- We use the daily position time series to **estimate the tectonic rate and the associated uncertainties** with the GMWMX-1 and the MLE.
- As each GNSS station records observations for the three coordinates (East, North, Up) and that the mean size of each time series is approximately 10 years, ranging from 8 to 15 years, the computing time for the GMWMX-1 for the whole GNSS network is below 40 seconds, while in comparison, Hector's processing time is approximately 23 minutes.

Case Study

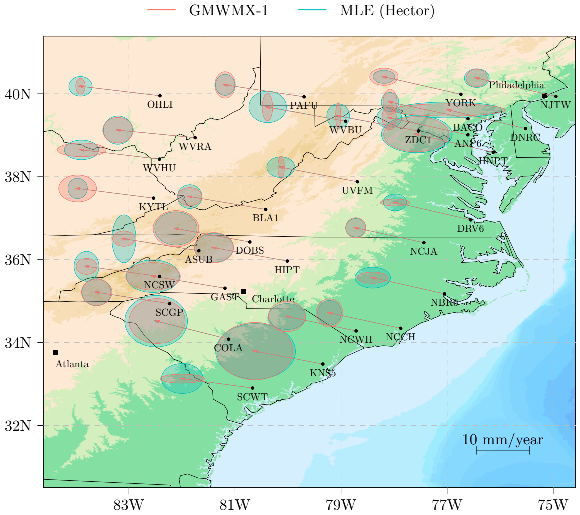


Figure: Estimated North-East velocity solutions for 33 GNSS receivers distributed over the East coast of the USA using the GMWMX-1 and Hector software (MLE).

Implementation

The `gmwmx` R package implements the Generalized Method of Wavelet Moments with eXogenous inputs estimator (GMWMX) and is available on CRAN and GitHub.

gmwmx 1.0.3 | Examples | Reference

`gmwmx` Overview

SMAC Group

The `gmwmx` `R` package implement the Generalized Method of Wavelet Moments with Exogenous Inputs estimator (GMWMX) introduced in Cucci, D. A., Voirol, L., Kermarrec, G., Montillet, J. P., and Guerrier, S. (2022) and provides functions to estimate times series models that can be expressed as linear models with correlated residuals. Moreover, the `gmwmx` package provides tools to compare and analyze estimated models and methods to easily compare results with the Maximum Likelihood Estimator (MLE) implemented in Hector, allowing to replicate the examples and simulations considered in Cucci, D. A., Voirol, L., Kermarrec, G., Montillet, J. P., and Guerrier, S. (2022). In particular, this package implements a statistical inference framework for the functional and stochastic parameters of models such as those used to model Global Navigation Satellite System (GNSS) observations, enabling the comparison of the proposed method to the standard MLE estimates implemented in Hector.

Find the package vignettes and user's manual at the package website.

Below are instructions on how to install and make use of the `gmwmx` package.

Installation Instructions

The `gmwmx` package is available on both CRAN and GitHub. The CRAN version is considered stable while the GitHub version is subject to modifications/updates which may lead to installation problems or broken functions. You can install the stable version of the `gmwmx` package with:

```
install.packages("gmwmx")
```

For users who are interested in having the latest developments, the GitHub version is ideal although more dependencies are required to run a stable version of the package. Most importantly, users **must** have a (`C++`) compiler installed on their machine that is compatible with R (e.g. `Clang`).

Links

View on CRAN

License

AGPL-3

Citation

Citing gmwmx

Developers

Davide Antonio Cucci
Author

Lionel Voirol
Author, maintainer

Stéphane Guerrier
Author

More about authors...

Dev status

CRAN 1.0.3
License GNU Affero General Public License v3.0
R>= 4.0.0
downloads 245/month
downloads 3119

Contributions & Further work

- We propose a **computationally efficient** and **scalable** estimator based on simple statistical concepts which allow to process large-scale networks which include thousands of GNSS stations. Our estimator is implemented in an open-source software available on CRAN.
- The first estimator (GMWMX-1) is highly computationally efficient but comes at the price of marginally deteriorated statistical properties. The second estimator (GMWMX-2) is asymptotically efficient for the linear functional parameters but has a slightly increased processing time.
- We are currently working on developing the theory to extend the GMWMX in a **robust setting**.
- The GMWMX estimator is well-suited for managing very large datasets, such as those encountered in air pollution studies (Chen and Zhou, 2020). In such cases, a prevalent tactic involves employing a **divide and conquer strategy** for estimation, as processing the complete dataset on a single server is unfeasible.

Thank You!

More info:

- Original article published in the *Journal of Geodesy*
- `https://lionelvoirol.com`
- Lionel.Voirol@unige.ch

References I

- Bos, M. S. et al. (2008). "Fast error analysis of continuous GPS observations". In: *Journal of Geodesy* 82.3, pp. 157–166. ISSN: 1432-1394. DOI: 10.1007/s00190-007-0165-x.
- Bos, Machiel S. et al. (2020). "Introduction to Geodetic Time Series Analysis". In: *Geodetic Time Series Analysis in Earth Sciences*. Cham: Springer International Publishing, pp. 29–52. DOI: 10.1007/978-3-030-21718-1_2.
- Bos, MS et al. (2013). "Fast Error Analysis of Continuous GNSS Observations with Missing Data". In: *Journal of Geodesy* 87.4, pp. 351–360.
- Chen, Lanjue and Yong Zhou (2020). "Quantile regression in big data: A divide and conquer based strategy". In: *Computational Statistics & Data Analysis* 144, p. 106892.
- Guerrier, Stéphane et al. (2013). "Wavelet-variance-based estimation for composite stochastic processes". In: *Journal of the American Statistical Association* 108.503, pp. 1021–1030.
- Guerrier, Stéphane et al. (2022). "Robust Two-step Wavelet-based Inference for Time Series Models". In: *Journal of the American Statistical Association*, pp. 1–18.
- He, X et al. (2019). "Investigation of the noise properties at low frequencies in long GNSS time series". In: *Journal of Geodesy* 93.9, pp. 1271–1282.

References II

He, Xiaoxing et al. (2017). "Review of current GPS methodologies for producing accurate time series and their error sources". In: *Journal of Geodynamics* 106, pp. 12–29.

He, Xiaoxing et al. (2021). "Spatial Variations of Stochastic Noise Properties in GPS Time Series". In: *Remote Sensing* 13.22, p. 4534. ISSN: 2072-4292. DOI: 10.3390/rs13224534.