Synthetic Sensor Data Generation Framework - Mathematics

Astrid Marie Skålvik

September 2025

This note contains a description of the synthetic data generation process in the R **synthsensor** package, available from https://github.com/AstridMarie2/synthsensor.

Parts of the mathematics are described in:

Astrid Marie Skålvik, Pekka Parviainen, Kjell-Eivind Frøysa, Ranveig N. Bjørk and Camilla Saetre, *A Bayesian Approach for Error Estimation and Uncertainty Assessment of Oceanographic Sensor Data*. Submitted (2025).

Contents

1 Background signal

The background signal or measurand, x_t can be modeled as a constant value, as an auto-regressive process (AR(1)), with random walk as a special case, a Poisson moving average or as a sine wave. For the Poisson moving average and the sine wave, the same background signal is used for both sensors. For the AR(1) and random walk, it is possible to generate two separate measurands $x_{1,t}$ and $x_{2,t}$, by adjusting a sensor cross-correlation factor.

1.1 Autoregressive process (AR(1))

For generating a stationary measurand x_t^{true} where the degree of temporal autocorrelation can be closely controlled, we use an autoregressive process:

$$x_t^{\text{true}} = \phi_x^{\text{true}} \cdot x_{t-1}^{\text{true}} + \varepsilon_{x,t}, \quad \varepsilon_{x,t} \sim \mathcal{N}(0, ({\sigma_x^{\text{true}}}^2)$$

Here ϕ_x^{true} is the autocorrelation coefficient, limited here between 0 and 0.99, and the noise term $\varepsilon_{x,t}$ is normally distributed with the innovation standard deviation σ_x^{true}.
For modeling situations where two sensors are measuring two slightly different measurands, for instance if the sensors are separated by a certain distance, it is possible to model two correlated measurands $x_t^{\text{true}(1)}$ and $x_t^{\text{true}(2)}$, by specifying a correlation coefficient $\rho_{1,2}^{\text{true}}$ and impose:

$$\text{Cov}(\epsilon_t^{(1)}, \epsilon_t^{(2)}) = \rho_{1,2}^{\text{true}} \cdot (\sigma_x^{\text{true}})^2$$

1.2 Random Walk

A random walk is a special case of the AR(1) process, where $\phi_x^{\text{true}} = 1$. This process generates a non-stationary measurand with strong temporal autocorrelation.

$$x_t = x_{t-1} + \epsilon_{x,t}, \quad \epsilon_{x,t} \sim \mathcal{N}(0, (\sigma_x^{\text{true}})^2)$$

Again, it is possible to generate two correlated measurands, as detailed for the AR(1) process above.

1.3 Poisson Moving Average

For generating a measurand x_t^{true} with a weak temporal autocorrelation through a smoothing process, we use the Poisson moving average defined as:

$$x_t^{\text{true}} = \frac{1}{k} \sum_{i=0}^{k-1} P_{t-i}, \quad P_t \sim \text{Poisson}(\lambda)$$

Here λ controls the noise level, and the degree of smoothing is controlled through the moving average window length k.

1.4 Sine Wave

For generating a measurand x_t^{true} with periodic variation:

$$x_t^{\text{true}} = A_x \cdot \sin\left(\frac{2\pi t}{T_{\text{sin}}}\right),$$

where A_x controls the amplitude and T_{sin} controls the length of the sine period.

2 Sensor delay and attenuation

Delay and attenuation in sensor response is modeled as an exponential moving average (EMA) of the measurand signal:

$$y_t^{(i)} = \alpha^{\text{synth}(i)} \cdot x_t^{\text{true}} + (1 - \alpha^{\text{synth}(i)}) \cdot y_{t-1}^{(i)}$$

Here $y_t^{(i)}$ denotes the simulated measurement data from sensor i at time t, before sensor noise, constant bias, spikes or drift errors are added.

3 Sensor noise and constant bias

A mean value μ_i is added to the background signal for sensor i. This allows for modeling situations where two sensors are highly correlated, with a fixed absolute deviation. Noise is added to sensor readings as normal distributions with specified standard deviations: $\varepsilon_y \sim \mathcal{N}(0, ({\sigma_y}^2)$. Correlation in sensor noise is specified through a correlation coefficient $\rho_{1,2}^{\text{noise}}$, so that $\text{Cov}(\epsilon_y^{(1)}, \epsilon_y^{(2)}) = \rho_{1,2}^{\text{noise}} \cdot \sigma_y^{(1)} \cdot \sigma_y^{(2)}$. This allows for modeling situations where some of the noise in sensor readings are due to common sources.

$$y_t^{(i)} \leftarrow y_t^{(i)} + \mu_i + \epsilon_y^{(i)}, \quad i = 1, 2$$

Here $y_t^{(i)}$ denotes the simulated measurement data from sensor i at time t, before spikes or drift errors are added. Anomaly flags for the whole timeseries is initialized to *Normal*.

4 Spikes in sensor readings

Spikes of random duration and magnitude are added to sensor readings at random time-steps. Both number of correlated spikes and uncorrelated spikes for each sensor, maximum spike duration and maximum spike amplitude can be adjusted. Correlated spikes are injected at the same time-step for both sensors, and uncorrelated spikes are injected randomly for each sensor time series.

Timing To avoid placing spikes too close to the edges of the time series, the spike start time $t_0^{(j)}$ of spike j is selected randomly from the interval

$$t_0^{(j)} \sim \text{DiscreteUniform}(100,\, n_{\text{timeseries}} - 100)$$

where $n_{\text{timeseries}}$ is the total length of the synthetic data time series.
The spike duration for spike j is noted $\Delta_{\text{spike}}^{(j)}$ and drawn from a discrete uniform distribution:

$$\Delta_{\text{spike}}^{(j)} \sim \text{DiscreteUniform}(\Delta_{\text{spike,min}}^{(i)}, \Delta_{\text{spike,max}}^{(i)})$$

Correlated spikes For each of the $n_{\text{spike,corr}}$ correlated spikes, spike amplitudes $A^{(1)j}$ and $A^{(2)j}$ are drawn from the same uniform distribution:

$$u^j \sim \text{Uniform}(0,\, 1)$$

$$A^{(i)j} = 2\sigma_y^{(i)} + u^j \cdot (\mu_y^{(i)} \cdot F_{\text{spike}} - 2\sigma_y^{(i)})$$

where $\mu_y^{(i)}$ is the sensor mean and $\sigma_y^{(i)}$ is the standard deviation of its normally distributed noise, and F_{spike} is a specified multiplier.

Uncorrelated spikes. Independently for each sensor i, for each of the $n_{\text{spike,uncorr}}^{(i)}$ uncorrelated spikes, spike amplitudes $A^{(i)j}$ are drawn from a uniform range:

$$A^{(i)j} \sim \text{Uniform}(2\sigma_y^{(i)},\, \mu_y^{(i)} \cdot F_{\text{spike}})$$

where $\mu_y^{(i)}$ is the sensor mean and $\sigma_y^{(i)}$ is the standard deviation of its normally distributed noise, and F_{spike} is a specified multiplier.

Spike injection Each spike j is injected at time steps $t \in [t_0^j,\, t_0^j + \Delta_{\text{spike}}^j - 1]$, with amplitude $A^{(i)j}$ and a random sign $D^{(i)j} \in \{-1, +1\}$ chosen with equal probability:

$$y_t^{(i)} \leftarrow y_t^{(i)} + D^{(i)j} \cdot A^{(i)j}, \quad i = 1, 2$$

The anomaly flag for this period is set to *Spike*.

5 Drifts in sensor readings

Drifts are simulated as gradual deviations added to the sensor readings over a specified duration.
The number of drifts is specified independently for each sensor. Ranges for duration and slopes are specified commonly for both sensors. All drifts are injected independently for each sensor, no correlated drifts are simulated.

Timing To avoid placing drifts too close to the edges of the time series, the drift start time $t_0^{(j)}$ is drawn from:

$$t_0^{(j)} \sim \text{DiscreteUniform}\left(100,\, n_{\text{ts}} - \Delta_{\text{drift}}^{(j)} - 100\right)$$

where $n_{\text{timeseries}}$ is the total length of the synthetic data time series. and $\Delta_{\text{drift}}^{(j)}$ is the duration of the drift, drawn from:

$$\Delta_{\text{drift}}^{(i)j} \sim \text{DiscreteUniform}\left(\Delta_{\text{drift,min}},\, \Delta_{\text{drift,max}}\right)$$

Slope For each drift event $j \in \{1 : n_{\text{drift}}^{(i)}\}$ on sensor i:
The drift slope is drawn from:

$$s^{(i)j} \sim \text{Uniform}\left(s_{\text{drift,min}},\, s_{\text{drift,max}}\right)$$

The range $[s_{\text{drift,min}},\, s_{\text{drift,max}}]$ can be chosen to also allow for drift in the negative direction.
The drift signal is modeled to increase linearly over its duration:

$$d_t^{(i)j} = \frac{k}{\Delta_{\text{drift}}^{(i)j} - 1} \cdot s^{(i)j}, \quad k = 0, 1, \ldots, \Delta_{\text{drift}}^{(i)j} - 1$$

Drift injection The drift $d_t^{(i)j}$ is added to the simulated sensor readings for time steps $t \in [t_0^{(j)},\, t_0^{(j)} + \Delta_{\text{drift}}^{(i)j} - 1]$:

$$y_t^{(i)} \leftarrow y_t^{(i)} + d_t^{(i)j}, \quad i = 1, 2$$

The anomaly flag for this period is set to *Drift*. If a spike is also present in the same time window, the anomaly flag is set to *Both*.