Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia
UNICO I+D
GOBIERNO DE ESPAÑA

SUCCESS-6G: EXTEND

WP3 Deliverable E7

Data-driven schemes for imperfect V2X communication

Project Title:	SUCCESS-6G: EXTEND
Title of Deliverable:	Data-driven schemes for imperfect V2X communication
Status-Version:	v1.0
Delivery Date:	14/02/2024
Contributors:	Charalampos Kalalas, Roshan Sedar, Pavol Mulinka (CTTC)
Lead editor:	Charalampos Kalalas (CTTC)
Reviewers:	-
Keywords:	Missing data; Data imputation; Expectation-maximization

Document revision history

Version	Date	Description of change
v0.1	12/12/2023	Table of contents defined
v0.2	15/01/2024	Initial content added
v0.3	31/01/2024	Additional content added
v1.0	14/02/2024	Final editions and version uploaded to the website

Disclaimer

Acknowledgment

The research conducted by SUCCESS-6G - TSI-063000-2021-39/40/41 receives funding from the Ministerio de Asuntos Económicos y Transformación Digital and the European Union-NextGenerationEU under the framework of the “Plan de Recuperación, Transformación y Resiliencia” and the “Mecanismo de Recuperación y Resiliencia”.

[1] http://creativecommons.org/licenses/by-nc-nd/3.0/deed.en_US

Executive Summary

Dependable measurement data are essential for the accuracy and integrity of vehicular state estimation by the maintenance center which performs condition monitoring tasks. However, vehicular networks are often subject to missing sensor observations due to -among others- channel stochasticity, hardware failures, and security attacks. In this deliverable, we study the problem of missing data in the vehicular measurement streams. We discuss the mechanisms that causally induce occlusions and investigate the ability of interpretable dynamical systems i) to fit the observed data at the aggregation point, and ii) to impute missing values by extracting knowledge from the spatiotemporal synergy among the ambient vehicular measurement space. A rigorous assessment of various missing data configurations based on empirical evaluations reveals meaningful performance trends for model fitting and recovery of incomplete information.

Table of Contents

List of Figures

List of Tables

1 Introduction

The acquisition of dependable measurement data is essential for the accuracy and integrity of vehicular state estimation by the maintenance center. Data aggregation points located at the network edge combine vehicular measurement trajectories captured at different locations and time instances to describe the evolution of vehicular state and model the rich interactions between quantities that co-evolve in time. However, vehicular networks are often subject to missing sensor observations due to the innate randomness of the wireless channel, hardware/equipment failures, security attacks, etc. Incompleteness in the aggregated data unavoidably affects the downstream processing tasks, leading to incomplete vehicular state knowledge posing risks in effective decision-making.

In this deliverable, we study the problem of missing data in the vehicular measurement streams. We discuss the mechanisms which causally induce occlusions (Section 2) and investigate the ability of interpretable dynamical systems i) to fit the observed data at the aggregation point, and ii) to impute missing values by extracting knowledge from the spatiotemporal synergy among the ambient vehicular measurement space (Section 3). A rigorous assessment of various missing data configurations based on empirical evaluations reveals meaningful performance trends for model fitting and recovery of incomplete information.

2 The problem of missing data

The aggregation of vehicular measurement streams constitutes an essential task in the value chain of vehicular networks and directly determines the integrity of the transmitted data and the resiliency of the acquisition infrastructure. Data aggregation points deployed at the edge combine measurement trajectories captured at different locations and time instances to describe the evolution of vehicular monitoring information to model the rich interactions between characteristics/variables that co-evolve in time [1] [2] [3].

Nevertheless, a key challenge for efficient vehicular data fusion and subsequent knowledge extraction resides in the completeness of aggregated information. In practice, the emergence of missing data in the fused vehicular measurement streams is inevitable. Missing information can be generally attributed to the following factors:

- Hardware failures: The malfunction of hardware components (e.g., synchronization failures or errors in sensor readings) may result in persistent missing observations for one or multiple state variables of the vehicle. In the case of interconnected in-vehicle systems, hardware failures may inadvertently occur in a cascade, where neighbouring sensors become progressively compromised in a short period. Cascade data occlusions with temporal dependency often become challenging to deal with, and they may hinder the effectiveness of reconstruction techniques.
- Connectivity issues: The imperfections of the underlying vehicular connectivity constitute an inseparable aspect of the data acquisition procedure. For example, the unreliable nature of the shared wireless medium may result in connectivity outages and packet losses for consecutive time-steps. The induced signal distortion leads to aggregated data inconsistencies and partial observability of the vehicular condition which, in turn, may adversely affect inference methods.
- Security attacks: The pervasive digitalization of vehicular systems introduces vulnerabilities and threat vectors, opening entirely new questions from a security and privacy perspective. Across all stages of the data acquisition chain, an increased number of entry points becomes available for potential adversaries to exploit and execute malicious attacks. For example, systematic modification of monitoring information and zero-injection measurements may perniciously mislead the monitoring operation of the vehicles.

3 Methods to recover/reconstruct missing information

3.1 Imputation problem as a dynamic Bayesian network

Dynamical systems offer an interpretable mathematical framework to i) learn the hidden patterns of time-series sensor data which exhibit high spatiotemporal correlation and ii) mine their underlying dynamics to gain insight into the evolution of the process being monitored. As such, dynamical systems provide an effective means for imputation of missing data and compression of the aggregated content at a fusion center.

At the aggregation point located at each vehicular edge node, the received measurements can be represented by a partially observable time sequence $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T]$, where each vector $\mathbf{y}_t$ contains the received measurements at time-step t from the deployed sensors. The stochastic nature of the wireless channel may result in a received vector $\mathbf{y}_t$ with intermittent measurements. We consider a time sequence of latent variables (i.e., hidden states) $\mathbf{Z} = [\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_T]$ to model the dynamics and the hidden patterns of the received measurements. We also introduce an indicator matrix, ϕ, for the missing measurements, i.e., $\phi_{t,k} = 0$ whenever the k-th sensor measurement in $\mathbf{y}_t$ is missing at time t; otherwise, $\phi_{t,k} = 1$. Let us also denote the observed part of $\mathbf{Y}$ as $\mathbf{Y}_r$ and the missing part as $\mathbf{Y}_m$. Following the rationale of linear dynamical systems [4], our model for the received measurements at the fusion center can be described by the following two equations:

$$\mathbf{z}_{t+1} = A\mathbf{z}_t + \mathbf{w}_t, \quad (1)$$
$$\mathbf{y}_t = C\mathbf{z}_t + \mathbf{v}_t. \quad (2)$$

To capture temporal correlation, we assume that the latent variables at each time tick depend linearly on the previous values via the linear state transition matrix A. At each time tick, the received vector y_t, including both observed and missing sensor measurements, is assumed to be a linear function of the latent variables z_t via the linear projection matrix C. This mapping implicitly captures the spatial correlation among the different sensor measurements [5]. Both hidden state evolution and received measurement processes are corrupted by zero-mean white Gaussian noise, $\mathbf{w}_t$ and v_t, with covariance matrices, Q and R, respectively. Further, $\mathbf{w}_t$ and $\mathbf{v}_t$ are assumed to be independent. The initial state z_0 of the latent variables is also a Gaussian random variable with mean π_1 and covariance V_1. Therefore, the parameter vector of our model is $\theta = (A, C, Q, R, \pi_1, V_1)$.

Based on Eqs. (1) and (2), we can express the conditional probabilities for the hidden state and the received sequence, respectively, as follows:

$$P(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \exp\left\{-\frac{1}{2}D(\mathbf{z}_t, A\mathbf{z}_{t-1}, Q)\right\}(2\pi)^{-\frac{\kappa_1}{2}}|Q|^{-\frac{1}{2}}, \quad (3)$$
$$P(\mathbf{y}_t \mid \mathbf{z}_t) = \exp\left\{-\frac{1}{2}D(\mathbf{y}_t, C\mathbf{z}_t, R)\right\}(2\pi)^{-\frac{\kappa_2}{2}}|R|^{-\frac{1}{2}}, \quad (4)$$

where $D(\boldsymbol{\omega}_t, \boldsymbol{\mu}_t, \Xi) = (\boldsymbol{\omega}_t - \boldsymbol{\mu}_t)'\Xi^{-1}(\omega_t - \boldsymbol{\mu}_t)$ denotes the square of the Mahalanobis distance of a vector ω_t with mean vector μ_t and covariance matrix Ξ.

Based on the Markov property implicit in the model, the factored representation of the joint probability distribution of Z and Y is given by

$$P(\mathbf{Z}, \mathbf{Y} \mid \theta) = P(\mathbf{z}_1)\prod_{t=2}^{T} P(\mathbf{z}_t \mid \mathbf{z}_{t-1})\prod_{t=1}^{T} P(\mathbf{y}_t \mid \mathbf{z}_t) \quad (5)$$

and the joint log-likelihood can be written as

$$\log P(\mathbf{Z}, \mathbf{Y} \mid \theta) = -\frac{1}{2}(D(\mathbf{z}_1, \pi_1, V_1) - \log|V_1| - T(\kappa_1 + \kappa_2)\log 2\pi) - \sum_{t=2}^{T}\left(\frac{1}{2}D(\mathbf{z}_t, A\mathbf{z}_{t-1}, Q)\right) - \frac{T-1}{2}\log|Q| - \sum_{t=1}^{T}\left(\frac{1}{2}D(\mathbf{y}_t, C\mathbf{z}_t, R)\right) - \frac{T}{2}\log|R|. \quad (6)$$

Given that the received sequence $\mathbf{Y}$ is characterized by intermittent measurements due to imperfect cellular connectivity, our goal is to maximize the conditional expectation of the received data log-likelihood, i.e.,

$$L(\theta) = E_{\mathbf{Y}_m, \mathbf{Z}|\mathbf{Y}_r, \phi}[\log P(\mathbf{Z}, \mathbf{Y} \mid \theta)]. \quad (7)$$

To that aim, we apply an iterative expectation maximization (EM) algorithm following a coordinate descent procedure [6]. We provide the details in the following subsection.

3.1.1 The EM algorithm

3.1.1.1 Overview

The EM algorithm is a general iterative algorithm for maximum likelihood estimation in incomplete-data problems [1]. The range of problems that can be addressed by EM is remarkably broad and includes maximum likelihood for problems not usually considered to involve missing data, such as variance-component estimation and factor analysis [7]. The EM algorithm formalizes a relatively old ad hoc idea for handling missing data: i) replace missing values by estimated values, ii) estimate parameters, iii) re-estimate the missing values assuming the new parameter estimates are correct., iv) re-estimate parameters, and so forth, iterating until apparent convergence. Each iteration of EM consists of an expectation step (E-step) and a maximization step (M-step). The M step is particularly simple to describe: perform maximum likelihood estimation of θ just as if there were no missing data, that is, as if they had been filled in. The E-step finds the conditional expectation of the missing data given the observed data and current estimated parameters, and then substitutes these expectations for the missing data. We provide the details in the following.

3.1.1.2 The E step and the M step of EM

The EM algorithm provides an iterative method for finding the maximum likelihood estimates of θ based on the observed measurements, $\mathbf{Y}_r$, by successively maximizing Eq. (7). The E-step of EM algorithm requires computing $L(\theta)$ in Eq. (7). Based on Eq. (6), this computation amounts to deriving the following three expectations:

$$\hat{\mathbf{z}}_t \equiv E[\mathbf{z}_t \mid \mathbf{Y}], \quad (8)$$
$$P_t \equiv E[\mathbf{z}_t \mathbf{z}_t' \mid \mathbf{Y}], \quad (9)$$
$$P_{t,t-1} \equiv E[\mathbf{z}_t \mathbf{z}_{t-1}' \mid \mathbf{Y}]. \quad (10)$$

Let $\mathbf{z}_t^T$ and V_t^T denote $E(\mathbf{z}_t \mid Y_1^T)$ and $\mathrm{Var}(\mathbf{z}_t \mid Y_1^T)$, respectively, for the subsequence of received measurements until time τ. Note that $\mathbf{z}_0^1 = \pi_1$ and $V_0^1 = V_1$. Let also θ be an initialization of the parameter vector. The conditional expectations in Eqs. (8)-(10) can be expressed as

$$\hat{\mathbf{z}}_t = \mathbf{z}_t^T, \quad (11)$$
$$P_t = V_t^T + \mathbf{z}_t^T \mathbf{z}_t^{T\prime}, \quad (12)$$
$$P_{t,t-1} = V_{t,t-1}^T + \mathbf{z}_t^T \mathbf{z}_{t-1}^T{}'. \quad (13)$$

and their computation can be decomposed into the following sets of forward and backward recursion:
i) Forward recursion:

$$\begin{aligned}
\mathbf{z}_t^{t-1} &= A\mathbf{z}_{t-1}^{t-1}, && (14)\\
V_t^{t-1} &= AV_{t-1}^{t-1}A' + Q, && (15)\\
K_t &= V_t^{t-1}C'(CV_t^{t-1}C' + R)^{-1}, && (16)\\
\mathbf{z}_t^{t} &= \mathbf{z}_t^{t-1} + K_t(\mathbf{y}_t - C\mathbf{z}_t^{t-1}), && (17)\\
V_t^{t} &= V_t^{t-1} - K_tCV_t^{t-1} && (18)
\end{aligned}$$

ii) Backward recursion:

$$\begin{aligned}
J_{t-1} &= V_{t-1}^{t-1}A'(V_t^{t-1})^{-1}, && (19)\\
\mathbf{z}_{t-1}^{T} &= \mathbf{z}_{t-1}^{t-1} + J_{t-1}(\mathbf{z}_t^{T} - A\mathbf{z}_{t-1}^{t-1}), && (20)\\
V_{t-1}^{T} &= V_{t-1}^{t-1} + J_{t-1}(V_t^{T} - V_t^{t-1})J_{t-1}', && (21)\\
V_{t-1,t-2}^{T} &= V_{t-1}^{t-1}J_{t-2}' + J_{t-1}\left(V_{t,t-1}^{T} - AV_{t-1}^{t-1}\right)J_{t-2}', && (22)
\end{aligned}$$

where Eq. (22) is initialized as $V_{T,T-1}^{T} = (I - K_TC)AV_{T-1}^{T-1}$.

After calculating the conditional expectations of the latent variables (E-step), the M-step re-estimates the parameter vector θ to be used in the E-step. To estimate $\theta = (A, C, Q, R, \pi_1, V_1)$, we take the respective partial derivative of Eq. (7), set to zero, and solve for the value of each respective parameter.

In particular, the updated parameters are computed as follows:

i) Projection matrix:

$$\begin{aligned}
\frac{\partial L}{\partial C} &= -\sum_{t=1}^{T} R^{-1}\mathbf{y}_t\hat{\mathbf{z}}_t' + \sum_{t=1}^{T} R^{-1}CP_t = 0,\\
C^{\text{new}} &= \left(\sum_{t=1}^{T}\mathbf{y}_t\hat{\mathbf{z}}_t'\right)\left(\sum_{t=1}^{T}P_t\right)^{-1}. \qquad (23)
\end{aligned}$$

ii) Measurement noise covariance:

$$\begin{aligned}
\frac{\partial L}{\partial R^{-1}} &= \frac{T}{2}R - \sum_{t=1}^{T}\left(\frac{1}{2}\mathbf{y}_t\mathbf{y}_t' - C\hat{\mathbf{z}}_t\mathbf{y}_t' + \frac{1}{2}CP_tC'\right) = 0,\\
R^{\text{new}} &= \frac{1}{T}\sum_{t=1}^{T}(\mathbf{y}_t\mathbf{y}_t' - C^{\text{new}}\hat{\mathbf{z}}_t\mathbf{y}_t'). \qquad (24)
\end{aligned}$$

iii) State transition matrix:

$$\begin{aligned}
\frac{\partial L}{\partial A} &= -\sum_{t=2}^{T} Q^{-1}P_{t,t-1} + \sum_{t=2}^{T} Q^{-1}AP_{t-1} = 0,\\
A^{\text{new}} &= \left(\sum_{t=2}^{T}P_{t,t-1}\right)\left(\sum_{t=2}^{T}P_{t-1}\right)^{-1}. \qquad (25)
\end{aligned}$$

iv) State noise covariance:

$$\frac{\partial L}{\partial Q^{-1}} = \frac{T-1}{2}Q - \frac{1}{2}\left(\sum_{t=2}^{T} P_t - A^{\text{new}} \sum_{t=2}^{T} P_{t-1,t}\right) = 0,$$
$$Q^{\text{new}} = \frac{1}{T-1}\left(\sum_{t=2}^{T} P_t - A^{\text{new}} \sum_{t=2}^{T} P_{t-1,t}\right). \quad (26)$$

v) Initial state mean:

$$\frac{\partial L}{\partial \pi_1} = V_1^{-1}(\hat{\mathbf{z}}_1 - \pi_1) = 0, \quad (27)$$
$$\pi_1^{\text{new}} = \hat{\mathbf{z}}_1.$$

vi) Initial state covariance:

$$\frac{\partial L}{\partial V_1^{-1}} = \frac{1}{2}V_1 - \frac{1}{2}(P_1 - \hat{\mathbf{z}}_1\pi_1' - \pi_1\hat{\mathbf{z}}_1' + \pi_1\pi_1') = 0, \quad (28)$$
$$V_1^{\text{new}} = P_1 - \hat{\mathbf{z}}_1\hat{\mathbf{z}}_1'.$$

Finally, using the Markov property, the missing sensor measurements $\mathbf{Y}_m$ can be computed from the estimation of the latent variables as

$$E[\mathbf{Y}_m \mid \mathbf{Y}_r, \mathbf{Z}; \theta] = C^{\text{new}}\, E\left[Z|_{(t,k)}, \phi_{t,k} = 0\right]. \quad (29)$$

The Eqs. (8)-(22) (E-step) and Eqs. (23)-(28) (M-step) complete one iteration of the EM algorithm; these equations are alternated repeatedly until the difference $L(\theta^{\text{new}}) - L\left(\theta^{\text{old}}\right)$ changes by an arbitrary small amount ϵ.

An alternative procedure can be followed based on Bayesian updates using sampling by setting conjugate prior distributions over all parameters [8]. This method provides the added benefit of uncertainty quantification based on computed position densities over the parameter space. The computation is carried out by Gibbs sampling, which constitutes an iterative Markov chain Monte Carlo (MCMC) scheme [9]. Missing values can be iteratively imputed by computing their conditional expectation with respect to the values of observed measurements, the posterior expectations of latent variables and the updated parameter values.

3.2 Performance assessment

3.2.1 Dataset description

The VeReMi dataset [10] includes 19 misbehaviour attack types and models two road traffic densities: high-density (37.03 Vehicles/km2) and low-density (16.36 Vehicles/km2). A log file per vehicle is generated which contains basic safety messages (BSM) transmitted by neighbouring vehicles over its entire trajectory. Each attack type dataset contains a ground truth file to record the observed behaviour of all participating vehicles. BSMs constitute a three-dimensional vector for position, speed, acceleration and heading angle features. Figure 1 depicts a raw sample of BSM data for a single vehicle. For subsequent imputation analysis, we have considered the log file for a single vehicle and kept only the genuine information by properly removing the misbehaving attack data, since the attack detection and classification are considered irrelevant tasks to our problem. Synthetic dropouts are then used to generate missing data by uniformly selecting space-time points for occlusion.

	type	sendTime	sender	senderPseudo	messageID	pos	spd	acl	hed
17	4	25210.186332	33	10332	21562	[1393.9276845310885, 1203.692849621629, 0.0]	[0.049400297340067005, -0.686074278731542, 0.0]	[0.166603725521922, -2.313798731172836, 0.0]	[0.063269582720791, -0.997996472890729 1, 0.0]
19	4	25210.436332	33	10332	21595	[1393.9276845310885, 1203.692849621629, 0.0]	[0.049400297340067005, -0.686074278731542, 0.0]	[0.166603725521922, -2.313798731172836, 0.0]	[0.063269582720791, -0.997996472890729 1, 0.0]
21	4	25210.686332	33	10332	21622	[1393.9276845310885, 1203.692849621629, 0.0]	[0.049400297340067005, -0.686074278731542, 0.0]	[0.166603725521922, -2.313798731172836, 0.0]	[0.063269582720791, -0.997996472890729 1, 0.0]
25	4	25210.936332	33	10332	21696	[1393.9276845310885, 1203.692849621629, 0.0]	[0.049400297340067005, -0.686074278731542, 0.0]	[0.166603725521922, -2.313798731172836, 0.0]	[0.063269582720791, -0.997996472890729 1, 0.0]
27	4	25211.186332	33	10332	30134	[1394.1720072035407, 1201.94700381985, 0.0]	[0.183983214273645, -2.555169745474803, 0.0]	[0.158360369223177, -2.199314281648395, 0.0]	[0.063269582720943, -0.99799647289072, 0.0]
...	...	...	...	...	...	...	...	...	...
7165	4	25368.936332	33	10332	715829	[127.9440058255349, 885.9631063084606, 0.0]	[-8.275102433876064, -0.48628168363595903, 0.0]	[4.492259181435928, 0.263999685982995, 0.0]	[-0.9703792835172421, 0.24158651891312902, 0.0]
7179	4	25369.186332	33	10332	716850	[122.01936718688863, 885.6254381946134, 0.0]	[-3.793979120628686, -0.21788847407861903, 0.0]	[4.49260375347267, 0.25802779102927603, 0.0]	[-0.9731995855753991, 0.22996209825940903, 0.0]
7192	4	25369.436332	33	10332	717806	[122.01936718688863, 885.6254381946134, 0.0]	[-3.793979120628686, -0.21788847407861903, 0.0]	[4.49260375347267, 0.25802779102927603, 0.0]	[-0.9731995855753991, 0.22996209825940903, 0.0]
7205	4	25369.686332	33	10332	718725	[122.01936718688863, 885.6254381946134, 0.0]	[-3.793979120628686, -0.21788847407861903, 0.0]	[4.49260375347267, 0.25802779102927603, 0.0]	[-0.9731995855753991, 0.22996209825940903, 0.0]
7228	4	25369.936332	33	10332	720286	[122.01936718688863, 885.6254381946134, 0.0]	[-3.793979120628686, -0.21788847407861903, 0.0]	[4.49260375347267, 0.25802779102927603, 0.0]	[-0.9731995855753991, 0.22996209825940903, 0.0]

640 rows × 9 columns

Figure 1: BSM data for a specific vehicle (ID:33) in VeReMi

3.2.2 Results

In this section, we aim to validate our proposed imputation method described in Section 3.1 against simulation results and provide a performance evaluation in terms of reconstruction error. In our proposed scheme, we initialize our estimated received time sequence $\hat{\mathbf{Y}}$ with $\mathbf{Y}_r$ while the missing sensor measurements are initially reconstructed by means of linear interpolation and then iteratively imputed as in Eq. (29). The process continues by updating the expectations of the latent variables based on the newly imputed values of the missing measurements until convergence. We further assume that the noise covariances in θ constitute diagonal matrices.

The effectiveness of reconstruction is evaluated in terms of the mean squared error (MSE), defined as the average of the squared differences between the real and reconstructed missing measurements, i.e.,

$$\text{MSE}\,(\mathbf{Y}, \Phi, \hat{\mathbf{Y}}) = \frac{1}{\sum_{t,k}\,(1-\phi_{t,k})} \sum_{t,k}\,(1-\phi_{t,k})\big(Y_{t,k} - \hat{Y}_{t,k}\big)^2.$$

To reduce random effects, we repeat each simulation 100 times and we report the average of the MSE.

Table 1 shows the imputation performance in terms of MSE for randomly missing values among measurement streams and time-steps. It can be observed that the proposed imputation method is capable of drawing insight from the received measurement values to make valid inferences for the missing data. Imputation performance expectedly registers a decline with rising percentage of missing entries in the aggregation point, albeit not at prohibitive levels.

	Missing values % (x100)									
	0.05	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**0.45**	**0.5**
MSE	0.036	0.039	0.04	0.042	0.043	0.045	0.046	0.048	0.051	0.055

Table 1: Imputation performance for varying percentage of missing measurements

In the following, we have conducted two types of experiments, investigating primarily the effect of ensuing occlusions: (a) we randomly opted for 1, 2, and 3 measurement streams to exhibit consecutive missing values of varying length starting at a random point in time (Tables 2-4, respectively); and (b) we opted for all measurement streams to report missing values for different scenarios of length of ensuing occlusions (Table 5).

	Occlusion length									
	10	20	30	40	50	60	70	80	90	100
MSE	0.0049	0.0052	0.0057	0.0064	0.0069	0.0076	0.0088	0.01	0.0144	0.0171

Table 2: Imputation performance for varying occlusion length with one missing measurement stream

	Occlusion length									
	10	20	30	40	50	60	70	80	90	100
MSE	0.011	0.0115	0.0118	0.0121	0.0126	0.013	0.0137	0.0144	0.0151	0.0178

Table 3: Imputation performance for varying occlusion length with two missing measurement streams

	Occlusion length									
	10	20	30	40	50	60	70	80	90	100
MSE	0.013	0.0144	0.0156	0.0164	0.0173	0.018	0.0199	0.0225	0.0263	0.0299

Table 4: Imputation performance for varying occlusion length with three missing measurement streams

	Occlusion length									
	10	20	30	40	50	60	70	80	90	100
MSE	0.0784	0.083	0.091	0.1089	0.133	0.188	0.224	0.252	0.391	0.54

Table 5: Imputation performance for varying occlusion length with all missing measurement streams

We observe that reconstruction of missing values in the case of experiment (a) does not suffer from the limitations imposed by potentially unfavourable network conditions causing occlusions. The imputation mechanism takes advantage of spatiotemporal correlations based on the received measurement streams which are consistent with the underlying physical process. Nevertheless, the recovery of missing vehicular information in experiment (b) suffers from the fact that in the absence of any measurement stream for an extended period of time there is non-existent spatial information in each time-step for the dynamical model to exploit. Moreover, the temporal information is limited to the relatively distant time-steps with recorded measurement values. Hence, the predicted values are mainly influenced by the outcome of linear interpolation from the initial stage of data imputation process, registering significant mismatch with respect to the ground truth values.

3.3 The potential of deep learning solutions

In the context of SUCCESS-6G, we also aim to summarize state-of-the-art guidelines for deep learning topology design and hyper parameter tuning in data imputation problems. We will implement a general approach in PyTorch framework and extend it to two implementations of (i) variational autoencoders and (ii) Generative Adversarial Imputation Networks (GAINs). We will define a set of scenarios reflecting possible real-world issues when the measurements are not being received or the sensors stop working. We will define a base classifier and compare its results on (i) the original dataset and (ii) imputed dataset in each scenario. The performance will be compared to the elementary

imputation techniques, e.g., rolling average/min/max/, etc. Regarding the applicability of variational autoencoders in data imputation tasks, we aim to extend the work of the authors in [11] by (i) proposing guidelines on deep net design, (ii) thorough comparison to other available methods and (iii) application of the method on the real-world datasets provided by IDNEO.

4 Conclusions

In this deliverable, we have explored the ability of dynamical systems to mine measurement streams under incomplete received trajectories. Using an open-source dataset, and creating synthetic dropouts, we have evaluated the reconstruction error for different missing data configurations. The proposed imputation method is capable of extracting knowledge from the spatiotemporal synergy among the respective trajectories to make valid inferences for the missing data. Imputation performance expectedly registers a decline with rising percentage of missing entries in the aggregation point, albeit not at prohibitive levels.

References

[1] L. Li, J. McCann, N. S. Pollard, and C. Faloutsos, "Dynammo: Mining and summarization of coevolving sequences with missing values," in Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 507–516. [Online]. Available: https://doi.org/10.1145/1557019.1557078

[2] N. Srebro and T. Jaakkola, "Weighted low-rank approximations," in Proceedings of the Twentieth International Conference on International Conference on Machine Learning, ser. ICML'03. AAAI Press, 2003, p. 720–727

[3] S. L. Brunton and J. N. Kutz, Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control. Cambridge University Press, 2019.

[4] R. H. Shumway and raD. S. Stoffer, Time Series Analysis and Its Applications (Springer Texts in Statistics). Berlin, Heidelberg: Springer-Verlag, 2005.

[5] G. Welch and G. Bishop, "An Introduction to the Kalman Filter," tech. rep., USA, 1995.

[6] C. Kalalas and J. Alonso-Zarate, "Sensor data reconstruction in industrial environments with cellular connectivity," in 2020 IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (IEEE PIMRC '20), August 2020

[7] S. Roweis and Z. Ghahramani, "A Unifying Review of Linear Gaussian Models," Neural Computation, vol. 11, no. 2, pp. 305–345, 1999.

[8] A. Wills, T. B. Schon, F. Lindsten, and B. Ninness, "Estimation of Linear Systems using a Gibbs Sampler," IFAC Proceedings Volumes, vol. 45, no. 16, pp. 203–208, 2012. 16th IFAC Symposium on System Identification.

[9] R. E. Kass, B. P. Carlin, A. Gelman, and R. M. Neal, "Markov Chain Monte Carlo in Practice: A Roundtable Discussion," The American Statistician, vol. 52, no. 2, pp. 93–100, 1998.

[10] J. Kamel, M. Wolf, R. W. van der Hei, A. Kaiser, P. Urien, and F. Kargl, "VeReMi Extension: A Dataset for Comparable Evaluation of Misbehavior Detection in VANETs," in 2020 IEEE International Conference on Communications, 2020, pp. 1–6.

[11] T. McCoy, S. Kroon, and L. Auret, "Variational autoencoders for missing data imputation with application to a simulated milling circuit," IFAC-PapersOnLine, vol. 51, no. 21, pp. 141–146, 2018.