METHOD ARTICLE

Spatial-temporal data analysis of digital twin [version 1; peer review: 1 approved, 2 approved with reservations]

Xing He¹, Qian Ai¹, Bo Pan², Lei Tang³, Robert Qiu¹

¹Shanghai Jiao Tong University, Shanghai, China
²State Grid Shanghai Municipal Electric Power Company, Shanghai, China
³State Grid Shanghai Pudong Electric Power Supply Company, Xi’an, China

Abstract

Background: Digital Twin (DT) has proven to be one of the most promising technologies for routine monitoring and management of complex systems with uncertainties.

Methods: Our work, which is mainly concerned with heterogeneous spatial-temporal data, focuses on exploring data utilization methodology in DT. The goal of this research is to summarize the best practices that make the spatial-temporal data analytically tractable in a systematic and quantifiable manner. Some methods are found to handle those data via jointly spatial-temporal analysis in a high-dimensional space effectively. We provide a concise yet comprehensive tutorial on spatial-temporal analysis considering data, theories, algorithms, indicators, and applications. The advantages of our spatial-temporal analysis are discussed, including model-free mode, solid theoretical foundation, and robustness against ubiquitous uncertainty and partial data error. Finally, we take the condition-based maintenance of a real digital substation in China as an example to verify our proposed spatial-temporal analysis mode.

Results: Our proposed spatial-temporal data analysis mode successfully turned raw chromatographic data, which are valueless in low-dimensional space, into an informative high-dimensional indicator. The designed high-dimensional indicator could capture the ‘insulation’ correlation among the sampling data over a long time span. Hence it is robust against external noise, and may support decision-making. This analysis is also adaptive to other daily spatial-temporal data in the same form.

Conclusions: This exploration and summary of spatial-temporal data analysis may benefit the fields of both engineering and data science.

Keywords
Digital twin, data-driven, uncertainty, spatial-temporal data, big data analytics
Corresponding author: Xing He (hexing_hx@126.com)

Author roles: He X: Conceptualization, Formal Analysis, Methodology, Writing – Original Draft Preparation; Ai Q: Conceptualization, Writing – Review & Editing; Pan B: Data Curation, Investigation; Tang L: Funding Acquisition, Project Administration; Qiu R: Formal Analysis, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the National Natural Science Foundation of China (51907121, 61871265) awarded to Xing HE and by the Foundation from State Grid Shanghai Pudong Electric Power Supply Company (SGSHPD00YJJS2106751) awarded to Qian Ai and Xing HE. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2022 He X et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: He X, Ai Q, Pan B et al. Spatial-temporal data analysis of digital twin [version 1; peer review: 1 approved, 2 approved with reservations] Digital Twin 2022, 2:7 https://doi.org/10.12688/digitaltwin.17446.1

First published: 19 Apr 2022, 2:7 https://doi.org/10.12688/digitaltwin.17446.1
I. Introduction
Digital twin (DT) is regarded as one of the most promising enabling technologies for realizing Industry 4.0. DT refers to the digital representation that mirrors a real entity or system. To date, DT has been successfully implemented in different fields, e.g., aerospace, smart factories, and energy systems. Our work conducts a study into DT by reviewing existing data utilization methodology and presenting the best practices. The work is mainly concerned with heterogeneous spatial-temporal data utilization.

Spatial-temporal data utilization falls within big data analytics (BDA) under the data-driven paradigm. Compared to the model-based paradigm, data-driven paradigm is more suitable for today’s engineering systems. Engineering systems nowadays are trending towards increased complexity, diversity, uncertainty, and coupling. For such a system, we can hardly build an accurate mechanism model or explain how the system works or what will happen next. Conversely, what we have at hand are the massive spatial-temporal data collected by advanced sensors, 5G, cloud, IoT (internet of things), etc. The rapid development of these data technologies, together with the recent progress in data science, spurs the emergence of DT. To seamlessly bridge the gap between the data resources and the engineering demand, there is a need for DT to harness the spatial-temporal data in a systematic and quantifiable manner.

A. Organization of this work
This article discusses emerging DT technology, with an emphasis on spatial-temporal data utilization.

Section II summarizes recent research on DT and extracts some fundamental characteristics and attributes of DT. Section III moves to the spatial-temporal data utilization and its methodology. Engineering data in spatial-temporal form are essentially intractable in analysis for most tools. By summarizing existing independent practices in a systematic and quantifiable manner, a novel mode is proposed that makes the spatial-temporal data analytically tractable. In our methodology, high-dimensional property and randomness are two issues to consider when facing spatial-temporal data. Section IV gives a discussion on these issues in detail. Some questions are raised first, followed by an overview of the underlying theories/tools and indicator systems that provide the solutions. Section V describes a real digital substation in China, and validates the proposed methodology using domain-specific applications.

II. Framework of dt
DT has already been successfully implemented in many fields, so behind that must be some common components. This section summarizes recent research on DT, and then extracts some common characteristics and attributes as its main components.

A. DT characteristics
According to 13, four characteristics of the up-to-date DT are listed as follows:

C1. Robustness of the model against diversity and uncertainty in a complex environment, with a focus on how it supports domain-specific outcomes;

C2. Employment of advanced data science, e.g., BDA (big data analytics) and AI (artificial intelligence);

C3. A link to the real world potentially in real time via data;

C4. The ability to interact with the physical counterpart, in order to keep consistency and to evaluate “what if” scenarios.

B. DT attributes
As the extension of DT characteristics, we summarize three DT attributes, data-driven mode, seamless interaction, and closed-loop feedback.

1) A1. Data-driven mode:
A data-driven model means that we consider the dataset to be a mine which supplies fuels to the DT engine. In particular, high-dimensional indicators, in the form of high-dimensional statistics or deep features, are extracted from the raw data.

Data-driven models have become a natural and stressing topic in energy systems since 2016. Data-driven approaches are also characterized by model-free—we no longer heavily rely on physical models. This attribute enables DT to handle the scenarios where the physical parameters are unreliable or even totally unavailable. Thus a quick start of DT program is feasible for a complex system, in which the behaviors and disciplines of system cell units are strongly coupling.

2) A2. Seamless interaction:
DT aims to create living digital simulation models that update and change as their physical counterparts change. DT continuously learns and updates itself from multiple sources to represent its near real-time (operating) status. The learning system, learns from itself, using (historical) sensor data that convey various aspects of its operating condition; from human experts, such as engineers with domain-specific knowledge through manual intervention; from other similar systems; from other similar fleets of systems; and from the larger system and environment of which it may be a part.

3) A3. Closed-loop feedback:
There are closed-loop feedbacks in both the physical space and the digital one. After the physical/virtual actuators execute commands, feedback of the system/DT is generated. It is worth mentioning that after the feedback flows back to the data center, it could sponsor a new round of closed-loop, and hence a succession of iterations is attainable. Through these iterations, a continuous optimization may be acquired in digital space.

III. Spatial-Temporal analysis and its methodology
This section, concerned with data utilization—or more specially, spatial-temporal data modelling and data analysis—provides an overview of the underlying theories/tools and indicator systems that have been previously described in the literature. DT aims to conduct joint spatial-temporal analysis on the data in existence in a systematic and flexible manner. For this purpose, data utilization methodology, rather than a particular data-driven algorithm/approach, should be focused on.

There has been growing interest in developing data analysis tools in engineering systems, especially ones in which massive
data are routinely generated and processed for various monitoring, control, inferential, and dispatch purposes. To make full benefit of these data, a series of fundamental questions on data utilization are raised:

Q1. What is the spatial-temporal data?
Q2. What is the goal of data mining?
Q3. How is the high-dimensional information from spatial-temporal data in general acquired?
Q4. Which tools are competent for data modelling?
Q5. Which tools are competent for data model analysis?
Q6. How is the gap between the information in high-dimensional space and the domain-specific meaning of the system in the real world bridged?
Q7. What is the property of high-dimensional indicators?
Q8. Can we simply regard those ubiquitous uncertainties as white noise? If not, how do we to model them?
Q9. Can we integrate diverse DT models? If yes, how?

These questions are beyond physical mechanism, causal relationship, and low-dimensional statistics.

A. Basics of spatial-temporal data, big data analytics, and high-dimensional information

This section addresses Questions Q1~Q2.

For the data utilization task, we deal with a large number of variables ($N$ variables, spatial space) simultaneously, and each variable ($i = 1, ..., N$) samples time-series for a given duration of observation ($T$ sampling times, temporal space). A classical statistic theory treats fixed $N$ only (often small, typically $N < 6$). This fixed (small) $N$ is called the low-dimensional regime$^{16}$. In practice, we are interested in the case that $N$ can vary arbitrarily in size compared with $T$ (often $T$ is large, typically $T > 60$, $N > 20$, $c = N/T > 0$)$^{19}$. This fundamental requirement is the primary driving force for us to study big data analytics.

Spatial-temporal data mining is expected to contribute some high-dimensional information, which is not obvious in the raw data, with domain-specific meaning attached. High-dimensional statistics, and deep features are two main types of high-dimensional information.

B. (Spatial-temporal) data utilization architecture and tools

When the input (data) and output (information) of the data mining are clear, we move to Questions Q3~Q5.

High-dimensional information can hardly be acquired form the spatial-temporal data$^{16}$. This phenomenon has greatly spurred the emergence of two fields in data science, AI (artificial intelligence) and BDA (big data analytics). For each field, we discuss a high-focus tool—1) Deep learning, which does well in massive data modelling, for AI$^{17}$, and 2) High-dimensional statistics, or more concretely, RMT (random matrix theory), which is good at data analytics, for BDA$^{46}$. Both of the tools involve a series of (high-dimensional) methods for the jointly spatial-temporal modelling and analysis, and have already made profound impacts on multi-fields.

1) Deep learning and its advantages:

Deep learning is the state-of-the-art data mining algorithm. It extracts high-level, complex abstractions (called deep features) via a hierarchical learning process via Equation (1)$^{18,19}$.

$$y = f(x) = f^L \left( W^L \cdot \ldots \cdot \left( W^2 \cdot f^1 (W^1 x + b^1) + b^2 \right) + b^3 \right) + b^4 \right)$$

For the deep learning model, its parameters can be adjusted adaptively with little knowledge about the physical mechanism or causal relationship—only labelled ground-truth data are needed. As a result, a deep learning model could be generalized to different cases or even different system without making significant modifications. For instance, we use CNN (Convolutional Neural Networks) for computer version system modelling$^{20}$, LSTM (Long Short-Term Memory) for prediction$^{21}$, and deep reinforcement learning for strategy optimization$^{22}$. In addition, the performance of the deep network model on generalization task could be evaluated by the test error quantitatively, so as to ensure its effectiveness during a domain-specific task.

2) Big data analytics, random matrix theory, and their advantages:

BDA concretely obtains the high-dimensional statistics through jointly temporal-spatial analysis. BDA is expected to gain some insights from matrix-based variables, such as eigenvalue, or the matrix variate itself$^{23}$. The matrix-based variables are the inherent statistical correlation among those variables in the $N \times T$ (large-dimensional) matrix discussed for Q1. High in dimensionality, rather than large in size, makes these matrix-based variables analytically intractable.

RMT is developed to deal with the data matrix rigorously. The goal of RMT is to understand the joint eigenvalue distribution as the statistic analytics from the data matrix in the asymptotic regime$^{24}$. For instance, LESs (Linear eigenvalue statistics)$^{25}$ of the matrix follow Gaussian distributions for very general conditions. The statistical properties of these LES variables are mostly derivable and provable. In this sense, RMT is fundamental in nature.

C. RMT-based statistics

1) M-P Law for spectral analysis:

RMT, as a statistical tool with profound theoretical basis, is adapted to multivariate analysis. This section is mainly concerned with M-P (Marchenko-Pastur) Law$^{25}$, a most fundamental law that can help model many intractable practical system, especially those with numerous variables.
RMT is mainly concerned with two ensemble random matrices—GUE (Gaussian unitary ensemble) and LUE (Laguerre unitary ensemble). LUE, which support different $N$ and $T$, is more flexible in practice.

$$ A = \begin{cases} \frac{1}{2}(Y + Y^H), & Y \in \mathbb{C}^{N \times N}, \text{GUE;} \\ \frac{1}{N} YY^H, & Y \in \mathbb{C}^{N \times T}, \text{LUE.} \end{cases} $$

(2)

where $Y$ is a GRM (Gaussian random matrix) whose elements are i.i.d. (independent identically distributed) standard normal random variables.

Let $p_A(x)$ be the theoretical spectral density of $A$, and define the theoretical spectral distribution $F_A(x)$:

$$ F_A(x) = \frac{1}{N} \sum_{i=1}^{N} I(\lambda_i \leq x), $$

(3)

where $A$ is GUE or LUE matrix, $I(\cdot)$ represents the event indicator function. We investigate the convergence rate of the expected ESD $\mathbb{E}(F_A(x))$ to Semicircle Law and M-P Law.

Let $g_A(x)$ and $G_A(x)$ denote the empirical spectral density and ESD (empirical spectral distribution) of $A$, and the Semicircle Law and M-P Law say:

$$ g_A(x) = \frac{1}{2\pi\sqrt{1-x^2}}, \quad x \in [-2, 2], \text{GUE}; $$

$$ g_A(x) = \frac{1}{2\pi|x(x-s_1)(s_2-x)|}, \quad x \in [s_1, s_2], \text{LUE}; $$

(4)

where $s_1 = (1 - \sqrt{c})^2$, $s_2 = (1 + \sqrt{c})^2$, and $c = N/T$.

$$ G_A(x) = \int_{-\infty}^{x} g_A(u) \, du. $$

(5)

2) Universality Principle of RMT:

Universality principle\textsuperscript{23} enables us to perform various hypothesis testing under the assumption that the matrix entries are not Gaussian distributed but use the same test statistics as in the Gaussian case. This is the very reason why RMT can handle many practical system. Numerous studies\textsuperscript{2,28} demonstrate that M-P Law is universally valid—the asymptotic results are remarkably accurate for engineering data with relatively moderate matrix sizes such as tens.

D. Linear eigenvalue statistics and its statistical properties

This part addresses Questions Q6–Q7.

Consider a random matrix $\Gamma \in \mathbb{R}^{N \times N}$, and $M$ is the covariance matrix $M = \frac{1}{N} \Gamma^H \Gamma$. The LES $\tau_c$ of $\Gamma$ is defined in 24,29 via the continuous test function $\varphi : \mathbb{C} \rightarrow \mathbb{C}$,

$$ \tau_c = \sum_{i=1}^{N} \varphi(\lambda_i) = \text{Tr}(\varphi(M)), $$

(6)

where the trace of the function of a random matrix is involved.

1) Law of Large Numbers:

The Law of Large Numbers tells us that $N^{1/2} \tau_c$ converges in probability to the limit

$$ \lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \varphi(\lambda_i) = \int \varphi(\lambda)\rho(\lambda)d\lambda, $$

(7)

where $\rho(\lambda)$ is the probability density function of the eigen-value $\lambda$, which is given in Equation (4). Therefore, from Equation (6) and Equation (7), we deduce that

$$ \mathbb{E}(\tau_c) = N \int \varphi(\lambda)\rho(\lambda)d\lambda, $$

(8)

2) Central Limit Theorem:

Central Limit Theorem (CLT), as the natural second step, aims to study the fluctuations of LES. The CLT for LES $\tau_c$ is stated as follows:

**Theorem III.1.** (M. Sheiberina, 2009,\textsuperscript{23}) Let the real valued test function $\varphi$ satisfy condition $\|\varphi\|_2 < \infty$ ($\varepsilon > 0$). Then $\tau_c$ as defined in Equation (6), in the limit $N, T \to \infty$, $c = N/T \leq 1$, converges in the distribution to the Gaussian random variable with the mean $\mathbb{E}(\tau_c)$, according to Equation (8), and the variance:

$$ \sigma^2(\tau_c) = \frac{2}{\pi^2} \int_{-\infty}^{\infty} \psi^2(\theta_1, \theta_2)(1 - \sin \theta_1 \sin \theta_2)d\theta_1d\theta_2 $$

$$ + \frac{k_i}{\pi^2} \left( \int_{-\infty}^{\infty} \varphi(\zeta(\theta)) \sin \theta \, d\theta \right)^2, $$

(9)

where $\psi(\theta_1, \theta_2) = \frac{\varphi(\zeta(\theta_1))}{\theta_1} \bigg|_{\theta_1 = \theta_2} = \frac{\varphi(\zeta(\theta))}{\theta_2} \bigg|_{\theta_2 = \zeta(\theta_1)}$,

and $\zeta(\theta) = 1 + c(1 + 2d \sin \theta)$; $k_i = \mathbb{E}(X_{ij}^2) - 3$ is the $4$th cumulant of entries of $X$.

3) LES-based hypothesis testing:

With the Law of Large Numbers and Central Limit Theorem, We formulate the hypothesis testing in terms of the statistical properties of LES. We can employ LES for anomaly detection formulating as:

$$ H_0 : \frac{\tau_c - \mathbb{E}(\tau_c)}{\sigma(\tau_c)} < \varepsilon, $$

$$ H_1 : \frac{\tau_c - \mathbb{E}(\tau_c)}{\sigma(\tau_c)} \geq \varepsilon, $$

(10)

where $\varepsilon$ is a threshold value that needs to be preset.

Our previous work shows that LES is robust against data errors (e.g., data loss, data out-of-synchronization\textsuperscript{i}) and insusceptible to (independent) random noises (not limited to white noises\textsuperscript{s}), which is not true to those low dimensional statistics such as mean and variance of any single variable. All of these
statistical properties make LES a good matrix-based variable for anomaly detection task.

Moreover, these LES indicators supply a multiple view angle to gain insight into the system, providing a much better way, compared to the classical one, to harness the spatial-temporal data. By comparing the experimental values $\rho_{b}$ (always fully data-driven) with ideal theoretical values $\rho_{e}$ (guaranteed by theorems), the complicated system is understood statistically.

In general, RMT supplies us a data-driven approach to high-dimensional information extraction via sampling spatial-temporal data. With a pure mathematical procedure, a cluster of statistical indicators is formed as a new epistemology for the system. Some advantages of this epistemology—such as data-driven and model-free mode, fast in speed, reasonableness, sensitivity, flexibility, and robustness against bad data—have already been shown in our previous work.

IV. RMT-Based modelling and analysis for Non-Gaussian Matrix-Variable

This section addresses Questions Q8~Q9.

The motivation behind this section arises from a key observation: The randomness in engineering often cannot be regarded as a white noise, and hence is often analytically intractable. This is especially true and important to a modern energy system where uncertainty is ubiquitous for almost every member. The entries of the observed matrix are random variables and large in size, so RMT is naturally relevant to the problem.

A. RMT-based factor analysis

1) Background of RMT-based factor analysis:

In practice, the collected data are oftentimes a mixture of the signals and the noise, and the noise is not simply i.i.d Gaussian matrix-variable. RMT-based FA (Factor Analysis) is employed to conduct dimension reduction of those high-dimensional datasets. This approach is capable of addressing the potential latent constructs (e.g., spatial-temporal independence) within the noise ingredient in the sampling data.

Question Q8 tells that assumptions and simplifications should be treated with caution, since we are facing high uncertainty aroused by the amount of diverse sources, although each source may only contribute a little. As the evidence that 31 tells that an accurate estimated value of regression-based problem is acquirable only if the noise is well addressed; even with a small error in measurements, the regression algorithms may fail in the task.

2) Formulation of factor analysis:

FA has already been successfully applied in various fields such as statistics and econometrics. Regarding the empirical data, FA is formulated as

$$X = LF + R,$$

where $X \in \mathbb{R}^{NT}$ is empirical data, $F \in \mathbb{R}^{pT}$ is a matrix of common factors, $L \in \mathbb{R}^{pN}$ is a matrix of factor loadings, $p$ is the number of factors, and $R \in \mathbb{R}^{NT}$ is a matrix of residues, also called unique factors or error variation.

Equation (11) provides us a way to decompose real-world data into systematic information and idiosyncratic noise. Usually, only $X$ is observable, $L$ is composed of the first $p$ principal components of $X$, $F = (L^\top L)^{-1}L^\top X$, and $R = X - LF$.

Then we define the empirical spectral density (ESD) of the covariance matrix of residues $R$ constructed from empirical data. It can be controlled by the $p$ number of common factors to be removed in Equation (11). It is defined as

$$\rho_{e}(\lambda) = \frac{1}{N} \sum_{i=1}^{N} \delta(\lambda - \lambda_{(i)})$$

(12)

where $\{\lambda_{(i)}\} = \{C_{N}^{1}\}_{i=1}^{N}$ is the eigenvalues of $C_{N} = \frac{1}{T} RR^{\top}$, and $\delta$ is the Dirac delta function.

3) Formulation of RMT-based FA:

Equation (11) is relevant to our RMT-based big data analytics. Taking advantage of high-dimensional statistics that was made analytically tractable only recently, we are able to conduct RMT-based FA. Ref. 36 employs the free random variables (FRV) calculus to calculate the ESD of the sample covariance for several VARMA-type (Vector autoregression–moving average) processes. This work extends the structure limitation of the ingredient from i.i.d Gaussian noise to ARMA process. The derivation is RMT-based and mathematically rigorous; the theoretical result is nicely matched against the spectra obtained via Monte Carlo simulations. Here, we give a brief conclusion about.

First, two assumptions are made:

I. The cross-correlations of $X$ are effectively eliminated by removing $p$ factors, i.e., $R$ has sufficiently negligible cross-correlation.

II. The auto-correlations of $X$ are exponentially decreasing, i.e., the residues $R$ can be modelled as an AR(1) process: $R_{i} = bR_{i-1} + \epsilon_{i}$.

Under these assumptions, we can conduct spectrum analysis of the simplified model, and thus $\rho_{e}(\lambda)$ is computable. The major implementations are briefly given as follows:

1. The mean spectral density can be derived from the Green’s function $G(z)$ by using the Sokhotsky’s formula:

$$\rho_{c}(\lambda) = \frac{1}{\pi} \lim_{\varepsilon \to 0} \text{Im} G(\lambda + i\varepsilon).$$

(13)

2. The Green’s function $G(z)$ can be obtained from the moments’ generating function $M(z)$:

$$G(z) = \frac{M(z) + 1}{z}, \quad |z| \neq 0.$$  

(14)

3. $M(z)$ can be found by solving the polynomial equation:

$$a^{4}c^{2}M^{4} + 2a^{2}c(-1 + b^{2})z + a^{2}c)M^{3} + ((1 - b^{2})z^{2} - 2a^{2}c(1 + b^{2})z + (c^{2} - 1)a^{4})M^{2} - 2a^{4}M - a^{4} = 0,$$

(15)

where $a = \sqrt{1 - b^{2}}$, and $c = \frac{N}{T}$. 

Page 6 of 16
It is worth mentioning that when $b = 0$, $R$ is a Gaussian random matrix and its spectral density follows $\rho_T(0)$. On the other side, M-P Law says that for a LUE matrix $\Gamma \in \mathbb{C}^{N \times T}$ ($c = N/T \leq 1$), its spectral density $g_{\Gamma}(x)$ does follow M-P Law. The two spectral densities should be equivalent, i.e. $\rho_T(0)$ is equivalent to $g_{\text{MP}}$. Figure 1 illustrates this deduction.

4) Roadmap of RMT-based FA:
We summaries the roadmap of RMT-based FA as Figure 2.

The spectrum of a covariance matrix typically consists of two parts: A few spikes/outliers and the bulk. The former represents common factors that mainly drive the features, and the latter represents unique factors or error variation that arise from idiosyncratic noise. Motivated by these two parts, we consider a minimum distance between two spectral densities—a theoretical one ($\rho_T$) from an ideal covariance structure model, and an empirical one ($\rho_E$) relevant to our observed time-series data.

Thus, the RMT-based FA is connected with a convex optimization problem. For convenience, the metric distance such as Jensen-Shannon divergence can be studied on the set of spectral distribution.

\[ \arg\min_{b, p} D(\rho_T(b), \rho_E(p)) = \sum_i \rho_T(a_i, b_i). \] (16)

where $D(a, b) = a \log a + b \log b - 2v \log \frac{a+b}{v}$ with $v = \frac{a+b}{2}$. The convex optimization can be readily calculated using modern software toolbox such as http://cvxr.com/cvx/download/CVX

Figure 1. Spectral Density Distributions of $\rho_T(b)$ (in Equation 13) and $g_{\text{MP}}$ (following M-P Law).

Figure 2. Roadmap of RMT-based Factor Analysis (RMT-random matrix theory, AR-auto regression, FA-factor analysis, $R_k$-residual matrix, $\rho_E$-empirical spectrum, $\rho_T$-theoretical spectrum, r.v.-random variables).
V. Use case

This section takes real digital substation in China as a case study. Our proposed methodology are also connected to several domain-specific applications.

A. A digital substation with spatial-temporal data: from digitization to informatization

We have taken a 35 kV substation in Shanghai as an example. It deploys 238 sensors to cover all aspects of its daily operation. There are also some inspection robots to supply (infrared) images of transformer, Switch cabinet, etc. The digitalization of this substation is depicted in Figure 3. Some of the sampling data in the digital space are listed as follows:

- Time-series users-monitoring data: active/reactive power, bus voltage magnitude/phase angel (three-phase), etc.
- Time-series devices-monitoring data: inside/outside temperature/humidity data, partial discharge (PD) data, etc.
- Static descriptive data: device manufacturer/batch/life-time, fault/maintenance record, tap change monitoring, (infrared) images, etc.

B. Informatization—taking switchgears as an example

We focus on switchgears, or more concretely, the health condition assessment and fault diagnosis of the switchgears. Most of the accidents involving switchgears are caused by insulation ageing, an event usually accompanied with PD phenomenon. For PD detection and PD source location, several types of sensors are deployed and we rearrange the collecting data (digitalization results) as follows. $X$—data collected by a single switchgear cabinet

1) PD Data: [transient earth voltage, ultrasound measuring, oil chromatographic]
2) Operation Data: [current, voltage]
3) Environment Data: [temperature, humidity]
4) Static Data: [device manufacturer/batch/life-time]
5) Label Data: [fault/maintenance record]

Conventional PD indicators, which are always model-based or handcrafted, are capable of handling an ideal or typical situation only. Traditional PD detection algorithm aims to design an indicator based on above measuring data independently, or based on the interpretation by experienced experts. These indicators may work pretty well in an ideal or typical situation; in practice, however, they are insufficient. For some very common situations—for instance, the start-up of air-conditioner in the substation or of load such as percussion drill, the passing away of mobile phone in the substation or high-speed trains—the threshold of the handcrafted indicator can be hardly set to balance the sensitivity and reliability.

For the above scenes, a basic and natural idea is that we use the data collected by all the switchgear cabinets or even the data from other substations. Starting from this idea, we list the high-dimensional data as follows.

$\Omega$—all the relevant data
1) Data from all the cabinets: $S_1 = \{X_1, X_2, \cdots\}$
2) Data from all the substation: $\Omega = \{S_1, S_2, \cdots\}$

$X$, $S$, and $\Omega$ are typical kinds of spatial-temporal dataset. Oriented to health condition assessment and fault diagnosis of switchgears, we can initiate an early sub-DT, and (high-dimensional) indicators are attainable as the analytics results. In addition, to the data structure in the form of set of spatial-temporal dataset ($\Omega$), several up-to-date technologies, such as Kronecker tensor-product $^{37}$, blocking calculation $^{38}$, are developed along this direction.

Insulation aging is one of the most common causes of accidents, and is usually accompanied with monitoring data changing. It is worth noting that, however, conventional (model-based or handcrafted) indicators are oftentimes insufficient for this task in practice, due to the complexity mechanism and procedure of insulation aging. On the other side, the threshold of low-dimensional statistics (such as mean, variance or maximum value) are hard to set—the sampled value is particularly vulnerable to external noise.

That is no longer a problem for high-dimensional indicator. It is believed that, from the perspective of statistic correlation in digital space, there must be some distinction (although maybe unknown) between the internal equipment condition and external noise. Because of the existence of this distinct

---

Figure 3. Digitalization of some Digital Substation in China (Main Page, Software: DT System, Ver 1.0).
correlation, we are capable of designing an indicator through spatial-temporal analysis. For instance, Figure 4 (see also Extended data\textsuperscript{39}) shows the high-dimensional indicator derived from oil chromatographic data of a transformer. The raw data was collected by the authors of this paper using the PMS (power production management system) of State Grid Zhejiang Electrical Power co. Ltd, where the authors have an ongoing research project. The data represents three years of typical spatial-temporal data, and can be used for hypothesis testing using Equation (10), for the purposes of creating a LES indicator for anomaly detection (green boxes in Figure 4b).

VI. Conclusion
This article, with an emphasis on data utilization methodology, provides a tutorial on digital twin (DT) in an energy system. Spatial-temporal analysis is an alternative to its prevailing model-based counterpart, and of significant importance in gaining insight into a modern system. With a high efficient data utilization manner, DT has some advantages. DT, with its quick start mechanism and data-driven mode attribute, is much more accessible than the traditional mode to engineering in practice. Besides, the performance of DT-, with a realtime interaction and closed-loop feedback attribute, can be improved by increasing data. Moreover, DT, with the help of a concise yet comprehensive data mining architecture, is friendly to spatial-temporal data surrounding data utilization Questions Q1 \~ Q7. The randomness in form of non-Gaussian data are also discussed in Questions Q8 \~ Q9, for the purpose of tying our architecture more closely to practical energy system.

DT itself is an emerging and promising technology for digitization-informatization in modern energy system. Along the virtual test function, there are also some directions about intellectualization. For instance, operational strategy optimization

![Normalized Raw Oil Chromatographic Data](image1)

(a) Normalized Raw Oil Chromatographic Data

![Mean and Standard Deviation of LES](image2)

(b) LES Indicator $\tau$

Figure 4. LES Indicator $\tau$ for Equipment Condition Assessment.
for all parties in a VPP (virtual power plant)\textsuperscript{46}, and MAS (multi-agent system)\textsuperscript{41} can be employed to enhance the interaction of each party. DT can also help with the operation, dispatch, management, and electricity market, making our DT a good reference for engineering. The tutorial provided in our article can be used for researchers and engineers working in energy domain. We hope it will be helpful for further investigation and construction of digitization, informatization, and intellectualization of energy systems.

VII. Data availability

Underlying data
No underlying data are associated with this article.

Extended data
IEEEDataPort: Chromatographic data. https://dx.doi.org/10.21227/xp4t-tx87\textsuperscript{90}.

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Acknowledgements
Thanks are due to F. Tao for assistance with the experiments and valuable discussion on digital twin theory.

References


Open Peer Review

Current Peer Review Status: ? ? ✓

Version 1

Reviewer Report 11 May 2022

https://doi.org/10.21956/digitaltwin.18721.r26940

© 2022 Oxoli D. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Daniele Oxoli
Department of Civil and Environmental Engineering, Politecnico di Milano, Milan, Italy

The article presents an up-to-date, literature review-based conceptualization of methodological requirements for the implementation of Digital Twins in energy systems. The content technically sounds and both language and paragraphing are good.

While the methods and applications in the energy systems domain are properly discussed, I believe there is a partial lack of supporting datasets/experiments and conceptual extensions to other (many) domains where spatio-temporal data analysis is used. This does not affect the validity of the paper. However, the selected title has a too general meaning to be justified with the actual content.

I suggest either expanding Section V with additional use cases / synthetic experiments (possibly connected to other domains that use spatio-temporal data analysis - e.g. Earth Observations, early-warning systems, etc.) or, at least, dedicate a paragraph to discuss the above within the conclusion (Section VI). Alternatively, please modify the title to better target the actual content of the article.

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Partly

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Yes

*Are the conclusions about the method and its performance adequately supported by the findings presented in the article?*

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** GIS, Earth Observation, Geostatistics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 06 May 2022

https://doi.org/10.21956/digitaltwin.18721.r26939

© 2022 Bozek P. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Pavol Bozek**

Faculty of Materials Science and Technology in Trnava, Institute of Production Technologies, Slovak University of Technology in Bratislava, Trnava, Slovakia

The paper has an original, scientific character, related to spatial-temporal data analysis of digital twin. In this work authors focuses on exploring data utilization methodology in DT. The work mainly concerned with heterogeneous spatial-temporal data. The content of the article is consistent with the scientific area of the journal Digital Twin. The subject raised by the authors is current and so far rarely noticed by other authors publishing in this area.

The issue described may in the future contribute to improving the efficiency of the automation and may benefit the fields of both engineering and data science.

For a better clarification, please edit your paper as follows:

1. Extend the text of manuscript (example introduction or conclusion) to concrete results in the world and in Europe. Improve the quality of the paper by presenting the results of publications of researchers and experts that are registered in the world databases (WoS). These are, for example: Chosen numerical algorithms for interval finite element analysis\(^1\) or Case study: Performance analysis and development of robotized screwing application with integrated vision sensing system for automotive industry\(^2\), thanks.

2. Figure 3 should be contrasting and readable.

3. Conclusions and future work should be extended to contain practical applications based on research described in this paper - expand references.
4. Highlight the course of dependencies/relations in figure No. 2 - unify font in the figure.

5. Modify the mathematical expression (formula) No: 1 and 2.
I recommend indexing the post after the proposed modifications.

References

Is the rationale for developing the new method (or application) clearly explained?
Yes

Is the description of the method technically sound?
Yes

Are sufficient details provided to allow replication of the method development and its use by others?
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?
Yes

*Competing Interests*: No competing interests were disclosed.

*Reviewer Expertise*: Automation, robotics, mechatronics, process control, mathematical analysis.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 03 May 2022

https://doi.org/10.21956/digitaltwin.18721.r26937

© 2022 Bao J. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Jinsong Bao
College of Mechanical Engineering, Donghua University, Shanghai, China

This manuscript provides a very interesting study of the application of spatio-temporal data analysis to digital twins that is indeed noteworthy. There are several issues that need to be modified and improved:

1. The temporality of the digital twin is obvious, however, in the spatiality issue, the paper needs to give a detailed discussion, while in considering the engineering characteristics of the spatio-temporal problem, the authors’ description needs to be enhanced.

2. The paper gives 9 fundamental questions for spatio-temporal analysis, how to divide its dimensions should be given, such as basic of data (Q1-Q2), it is suggested to give the framework with a big picture.

3. The manuscript starts ambitiously, but ends abruptly, with a slight focus on details rather than methodology in the specific implementation techniques.

4. Paper citations are not good in terms of timeliness and more recent papers should be considered.

5. The language needs some improvement.

Is the rationale for developing the new method (or application) clearly explained?  
Yes

Is the description of the method technically sound?  
Yes

Are sufficient details provided to allow replication of the method development and its use by others?  
Yes

If any results are presented, are all the source data underlying the results available to ensure full reproducibility?  
Partly

Are the conclusions about the method and its performance adequately supported by the findings presented in the article?  
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Digital twin

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have
significant reservations, as outlined above.