ITISE 2022 (8th International Conference on Time Series and Forecasting)

SYNTHETIC SUBJECT GENERATION WITH COUPLED COHERENT TIME SERIES DATA

Xabat Larrea[1,2], Mikel Hernandez[1], Gorka Epelde[1,3], Andoni Beristain[1,3], Cristina Molina[1], Ane Alberdi[2], Debbie Rankin[4] , Panagiotis Bamidis[5], Evdokimos Konstantinidis[5,6]

[1] *Vicomtech Foundation*

[2] *Mondragon Unibertsitatea*

[3] *Biodonostia Health Research Institute*

[4] *Ulster University*

[5] *Aristotle University of Thessaloniki*

[6] *European Network of Living Labs*

Speaker: Mikel Hernandez

mhernandez@vicomtech.org

OUTLINE

1. INTRODUCTION
2. RELATED WORK
3. PROPOSED APPROACHES
4. IMPLEMENTATION
5. CONCLUSION

1. INTRODUCTION | 2. RELATED WORK | 3. PROPOSED APPROACHES | 4. IMPLEMENTATION | 5. CONCLUSION

MOTIVATION

H2020 VITALISE PROJECT

Provide virtual transnational access to data generated from several Living Labs throughout Europe and beyond.

PROBLEM

- A high amount of sensor data is generated daily and stored in Living Labs.
- This data cannot be shared externally as they contain personal and sensitive information.

SOLUTION

Incorporate SDG[1] techniques into a controlled data processing workflow [1] to generate shareable data for external researchers in compliance with the GDPR[2] .

[1] **SDG**: *Synthetic Data Generation*
[2] **GDPR**: *General Data Protection and Regulation*

[1] Hernandez, M., Epelde, G., Beristain, A., Álvarez, R., Molina, C., Larrea, X., Alberdi, A., Timoleon, M., Bamidis, P., Konstantinidis, E.: *Incorporation of Synthetic Data Generation Techniques within a Controlled Data Processing Workflow in the Health and Wellbeing Domain*. Electronics. 11, 812 (2022). https://doi.org/10.3390/electronics11050812

SYNTHETIC TIME SERIES GENERATION

WHY SYNTHETIC TIME SERIES?

- Data generated from Living Labs is mainly sensor data from subjects, usually treated as time series.
    - ✓ *Physiological data*
    - ✓ *Mobility track data*
- Traditional anonymization techniques tend to suppress useful data.
    - ✓ *Noise addition*
    - ✓ *Attributes deletion*

OBJECTIVE OF SYNTHETIC TIME SERIES

Make Living Labs data accessible for external researchers to conduct and validate experiments with GDPR[1] compliance.

- ✓ Ensure the *Privacy* of synthetic time series.
- ✓ Prioritize the *Utility* of synthetic time series over the *Resemblance* with the real time series.

[1] **GDPR**: *General Data Protection and Regulation*

SYNTHETIC TIME SERIES GENERATION APPROACHES

Supervised GAN

TimeGAN

NeuralProphet

Probabilistic AutoRegressive (PAR)

Conditional Temporal GAN (CTGAN)

Transformer-based Time-Series GAN (TTS-GAN)

- Most of them are focused on generating highly realistic sequences of data.
  - ✓ Do not consider the metadata of the subjects.
- Some of them transform time series data into the latent space.
  - ✓ Do not transform the generated synthetic time series back to the data format of the real time series.

GAN: *Generative Adversarial Network*

GENERATION OF SYNTHETIC SUBJECTS CONTAINING TEMPORAL DATA

A1: SYNTHETIC SUBJECTS WITH REAL TIME SERIES DATA

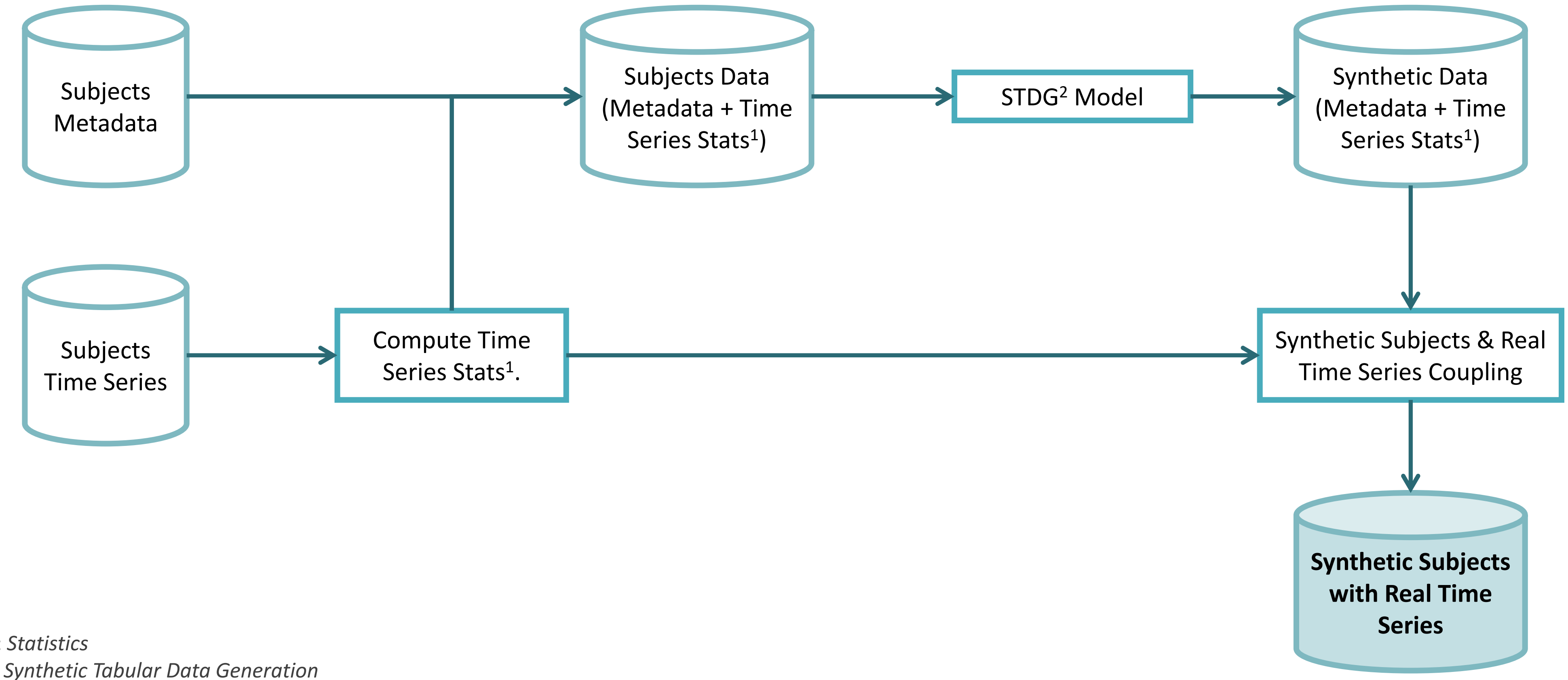
- Time series data does not require further transformations to ensure patients privacy.
- Synthetic subjects are generated and joined with real time series data.

A2: SYNTHETIC SUBJECTS WITH SYNTHETIC TIME SERIES DATA

- Time series data needs to be synthetically generated to ensure patients privacy.
- Well-performing STSG[1] techniques that consider the metadata of subjects are required.

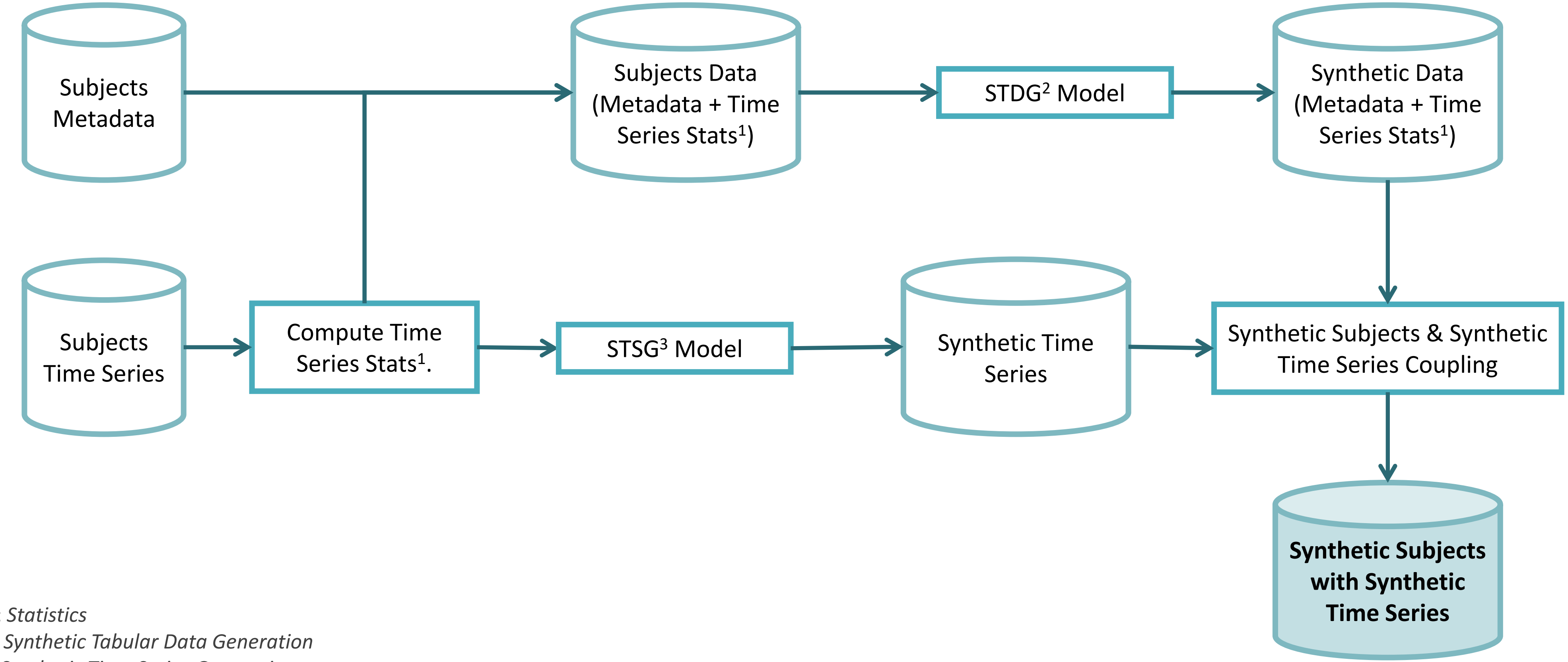
[1] **STSG**: *Synthetic Time Series Generation*

A1: SYNTHETIC SUBJECTS WITH REAL TIME SERIES DATA



[1] **Stats.**: *Statistics*
[2] **STDG**: *Synthetic Tabular Data Generation*

A2: SYNTHETIC SUBJECTS WITH SYNTHETIC TIME SERIES DATA



[1] **Stats.**: *Statistics*
[2] **STDG**: *Synthetic Tabular Data Generation*
[3] **STSG**: *Synthetic Time Series Generation*

IMPLEMENTATION STEPS FOR A1

1. DATASET SELECTION

The Treadmill Maximal Exercise Tests (TMET) Database from PhysioNet

2. EDA AND DATA PREPROCESSING

Data cleaning from 992 to 950 subjects.

3. TIME SERIES FEATURES EXTRACTION

- Maximal speed
- Test duration
- Max, min and mean values of the physiological variables

4. SYNTHETIC SUBJECT GENERATION

Synthetic subjects generation with metadata and time series features.

- TVAE[1] model from SDV[2]

5. TIME SERIES COUPLING WITH SYNTHETIC SUBJECTS

- Normalization of synthetic and real features.
- Finding the minimum Euclidean distance between features.

6. VALIDATION OF SYNTHETIC SUBJECTS

- Descriptive statistics (mean & std[3])
- PCA[4] comparison
- Correlation Matrices comparison
- Hamming distances

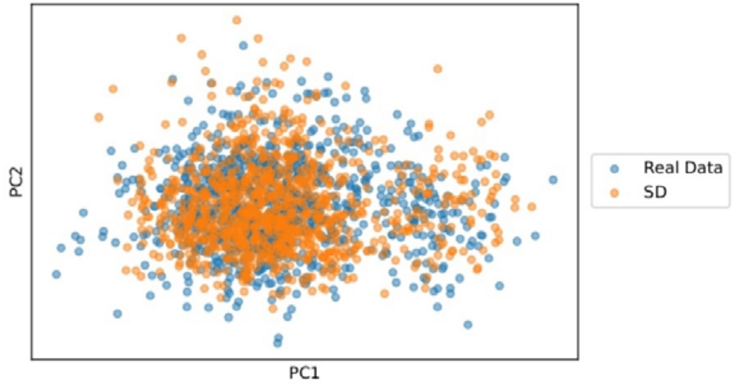
[1] **TVAE.**: *Tabular Variational Autoencoder*
[2] **SDV**: *Synthetic Data Vault*
[3] **std.**: *Standard Deviation*
[4] **PCA**: *Principal Component Analysis*

RESULTS

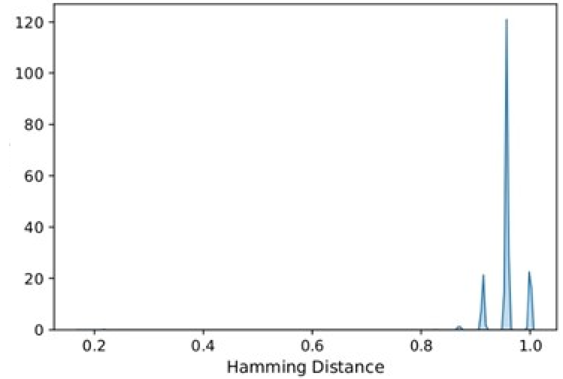
Descriptive Statistics

Variable	Real Data	Synthetic Data
Age	28.95 ± 10.19	27.67 ± 9.94
Weight	73.14 ± 11.96	72.12 ± 11.56
Height	174.82 ± 7.99	174.39 ± 7.73
Humidity	48.14 ± 8.54	45.4 ± 6.86
Temperature	22.82 ± 2.79	23.92 ± 1.5
Sex	Male (n = 806) Female (n = 104)	Male (n = 810) Female (n = 110)

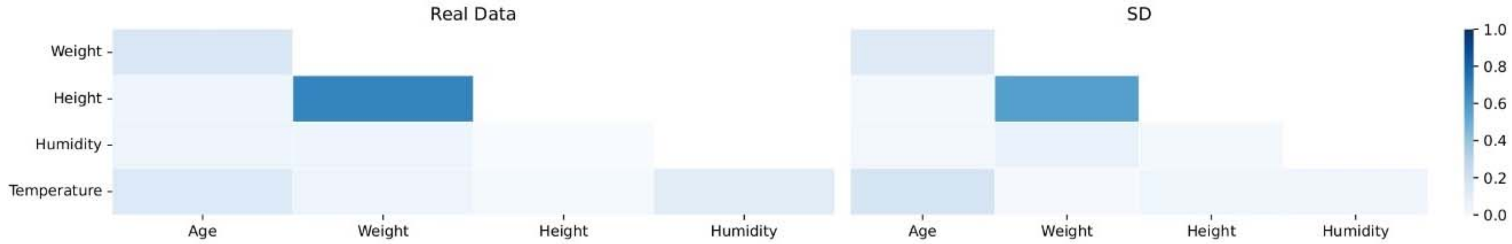
PCA[1] Comparison



Hamming Distance



Correlations Comparison



[1] **PCA**: *Principal Component Analysis*

MAIN CONCLUSION

The proposed approach for the generation of synthetic subjects with real time series data can be used to generate a synthetic, thus shareable dataset.

- With the selected dataset the proposed methodology can be used to generate synthetic patients and then combine them with real time series data.
- The generated cohort of synthetic subjects preserves the privacy of the cohort of real subjects while maintaining correlations and dimensional properties.

LIMITATIONS

- Lack of evaluation with other time series datasets to validate the generalizability of the implemented approach.
- Second proposed approach not implemented as a STSG[1] model that resembles similarity to real time series has not been found.
- Privacy of the generated synthetic subjects and the quality of the time series coupling have not been extensively analyzed.

[1] **STSG**: *Synthetic Time Series Generation*

FUTURE WORK

- Incorporate other missing time series data imputation techniques (e.g. forecasting) into the data processing step.
- Design a strategy to evaluate the quality of the synthetic subjects and real time series coupling based on forecasting analysis models.
- Evaluate the implemented approach with more time series datasets.
- Implement and validate the proposed second approach together with more STSG[1] techniques.
    - Define a strategy to validate and evaluate the temporal nature of the generated synthetic time series.
- Define a complete strategy to evaluate the subject's resemblance and privacy.
- Incorporate the proposed approaches in the "*VITALISE controlled data processing workflow to enable open access for research data processing*" presented in [1].

[1] **STSG**: *Synthetic Time Series Generation*

[1] Hernandez, M., Epelde, G., Beristain, A., Álvarez, R., Molina, C., Larrea, X., Alberdi, A., Timoleon, M., Bamidis, P., Konstantinidis, E.: *Incorporation of Synthetic Data Generation Techniques within a Controlled Data Processing Workflow in the Health and Wellbeing Domain*. Electronics. 11, 812 (2022). https://doi.org/10.3390/electronics11050812

ITISE 2022 (8th International Conference on Time Series and Forecasting)

VITALISE

Virtual Health & Wellbeing Living Lab Infrastructure

vicomtech

MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE

THANK YOU!

vitalise-project.eu

www.vicomtech.org

mhernandez@vicomtech.org

Speaker: Mikel Hernandez