SplitVAES - Decentralized scenario generation from siloed data for stochastic optimization problems

Islam H M Mohaimanul; Huynh Q. N. Vo; Paritosh Ramanan
School of Industrial Engineering and Management, Oklahoma State University

MOTIVATION

- Stochastic planning and optimization (SO) in **large-scale, multi-stakeholder, networked** infrastructure systems requires spatiotemporally interdependent data-driven scenarios
- Centralized aggregation of stakeholder data is challenging due to **privacy, computational, and logistical** bottlenecks.
- Spatiotemporal interdependencies are difficult to capture without centralized data aggregation from all stakeholders.

$$\min_{x} \; c^T x + \mathbb{E}_{s\in\mathcal{S}}\Big[\min_{y} \; q_s^T y_s\Big]$$
$$\text{s.t. } Ax = b$$
$$W_s y_s = h_s - T_s x \;, \forall s \in \mathcal{S}$$
$$x \geq 0 \text{ and } y_s \geq 0 \;, \forall s \in \mathcal{S}$$

Simple two-stage, linear, stochastic optimization formulation as an example

CONTRIBUTIONS

We propose a decentralized machine learning (ML) approach that:

- Leverages edge-based autoencoders with server-driven variational autoencoder **to capture global spatiotemporal interdependencies from siloed data**.
- Decomposes global backpropagation steps across edge and server models, **enabling the bi-directional flow of learning insights without moving raw data**.
- Enables **a scalable, real-world implementation to generate scenarios** for stochastic optimization problems without the need to move data.
- Handles **heterogeneous datasets across diverse areas** by comparing the generated scenarios with established benchmark methods.

FRAMEWORK DESIGN

- **Server-level variational autoencoder** (Server-VAE): yields **latent space** embeddings with the help of a probabilistic encoder, then uses the mean and variance of the latent embeddings to generate latent space representations; finally, the representations are fed to a probabilistic decoder to obtain the reconstructions of the raw data.
- **Edge-level autoencoder** (Edge-AE): deep neural network (DNN) encoder and decoder models; the encoders receive siloed time-series data from each stakeholder to yield low-dimensional embeddings, while the decoder reconstructs the input data from these embeddings.
- **Loss functions:** two loss functions are used in the training mechanism of our proposed framework.
  - Reconstruction Loss: the **binary cross-entropy function** is employed to measure the mean reconstruction error between predicted values and observed values.
  - Kullback-Leibler (KL) Loss: the **KL loss** measures the disparity between the learnt latent space embedding and the reference distributions.

$$L = -\frac{1}{N}\left[\sum_{j=1}^{N}\left[t_j \log(p_j) + (1-t_j)\log(1-p_j)\right]\right]$$

Binary cross-entropy as Reconstruction Loss

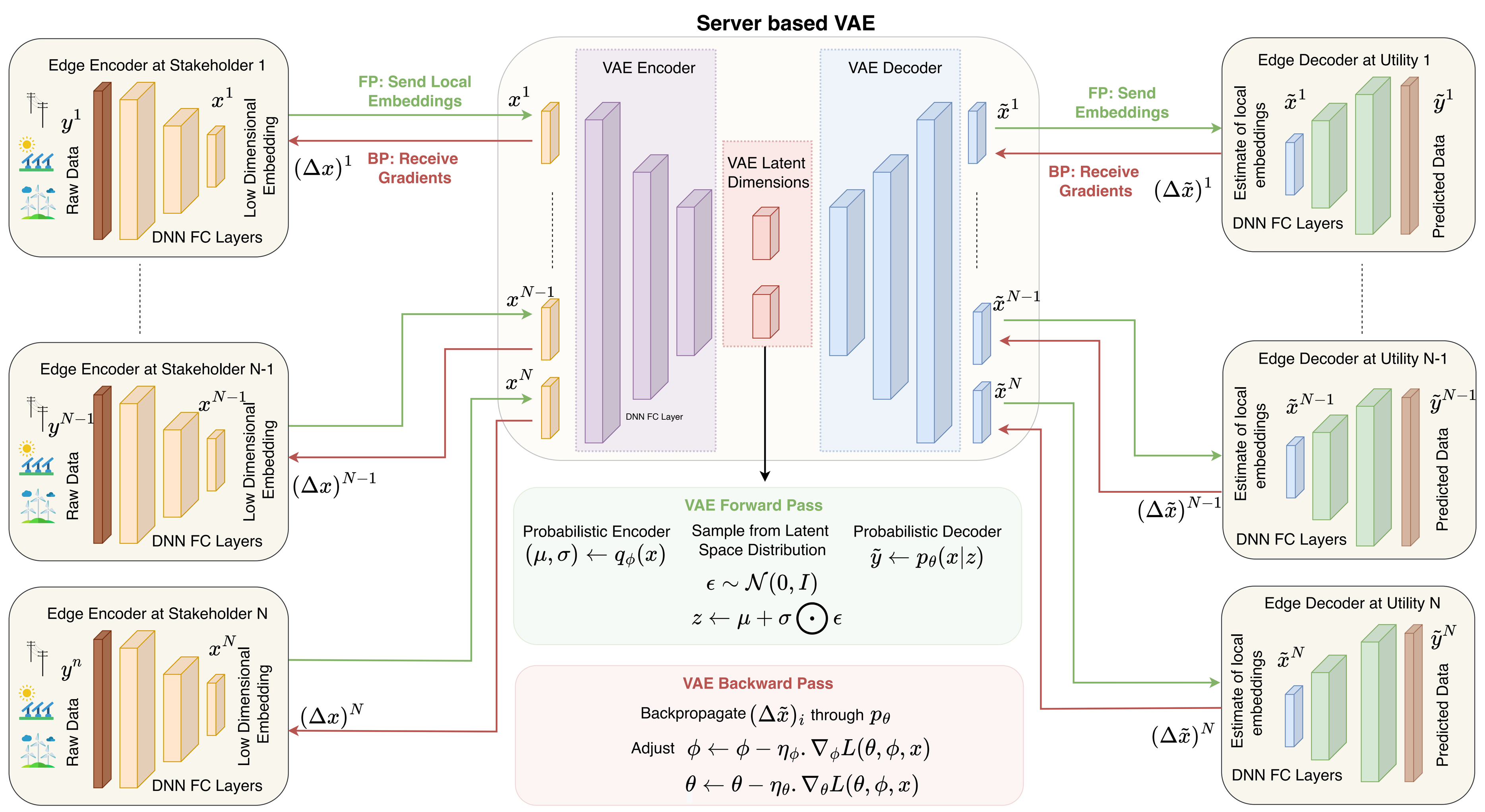
$$D_{\mathrm{KL}}(P \parallel Q) = -\sum_{x\in\mathcal{X}} P(x) \, \log\left(\frac{Q(x)}{P(x)}\right)$$

KL Divergence to compare two distributions

EVALUATION METHODS

- **Benchmark**: We compare the SplitVAEs with centralized scenario generation methods, including the **Gaussian copula** and the **Central-VAE** method. We also employ the **t-SNE technique** to visualize the kernel density distribution of the generated scenarios, transformed into one-dimensional embeddings.
- **Metrics**: We employ multiple quantitative evaluation metrics to measure the quality of the generated scenarios namely **Fréchet inception distance** (FID), **Energy Score** (ES); **Root-mean-square Error** (RMSE), and **Continuous Ranked Probability Score** (CRPS).
- **Datasets**: We primarily evaluated the decentralized scenario generation capability of our proposed framework using the **USAID**, **ACES**, and **ACTIVSg** datasets.
  - USAID: the Order Cycle Time information of **85 distributors from 2017 to 2023** are used.
  - ACES: carbon emission profiles for **25 petroleum refineries in Texas for summer seasons** in 2017.
  - ACTIVSg: **load** and **renewable data** from the 2000 bus transmission system test case for Texas.
- **Hardware Setup**: Our experiments are conducted on **Pete Supercomputer**.
- **Software Setup**: All experiments were conducted **OpenMPI** distributed memory framework, available as mpi4py, in Python. The **PyTorch** library is used for constructing, training and evaluating the machine learning models.

COMPUTATIONAL ARCHITECTURE



Forward Pass (FP):

- Edge-AE generates **low dimensional embeddings** which are sent to the Server-VAE.
- The probabilistic encoder $q_\phi(x)$ yields **latent space representations** of the received embeddings.
- The probabilistic decoder $p_\theta(x|z)$ generates **estimates of low dimensional embeddings** for each edge location.
- The Edge-AE decoder uses the Server-VAE embedding estimates to predict reconstruction error of raw data.

Backward Pass (FP):

- Edge-AEs receive embedding estimates and compute prediction error using decoder.
- The errors on the input layer are distributed back to the Server-VAE.
- The received errors are backpropagated through the VAE to yield errors at the VAE input layer.
- VAE input errors are distributed to Edge-AEs where they are backpropagated through the encoder

COMPUTATIONAL PERSPECTIVES AND BENEFITS

Planning and Optimization for Nodal Authorities

- SplitVAEs enable nodal authorities (NA) such as control towers in supply chains, and Independent System Operators (ISOs) in power systems to optimize and **plan for the entire network without needing to move stakeholder data.**
- Using SplitVAEs as a NA can generate scenarios by using the latent space dimensions learned by the Server-VAE in conjunction with the trained Edge-AEs to use for SO problems.

Applicability to a Diverse Case Studies:

- SplitVAEs performance is **nearly identical** to that of most centralized training schemes like the Gaussian copula and the Central-VAE in respect to evaluation metrics.
- All methods **can capture the underlying data distributions**, as indicated by the embedding distribution graphs and the spatiotemporal information of the observed data is captured in **a high-fidelity fashion**.

Handling Dimensional Heterogeneity

- The SplitVAEs model **can handle heterogeneity in embedding dimensions** as illustrated by varying the region decompositions in ACTIVSg dataset leading to different embedding sizes for each Edge-AE.

Architecture Scalability

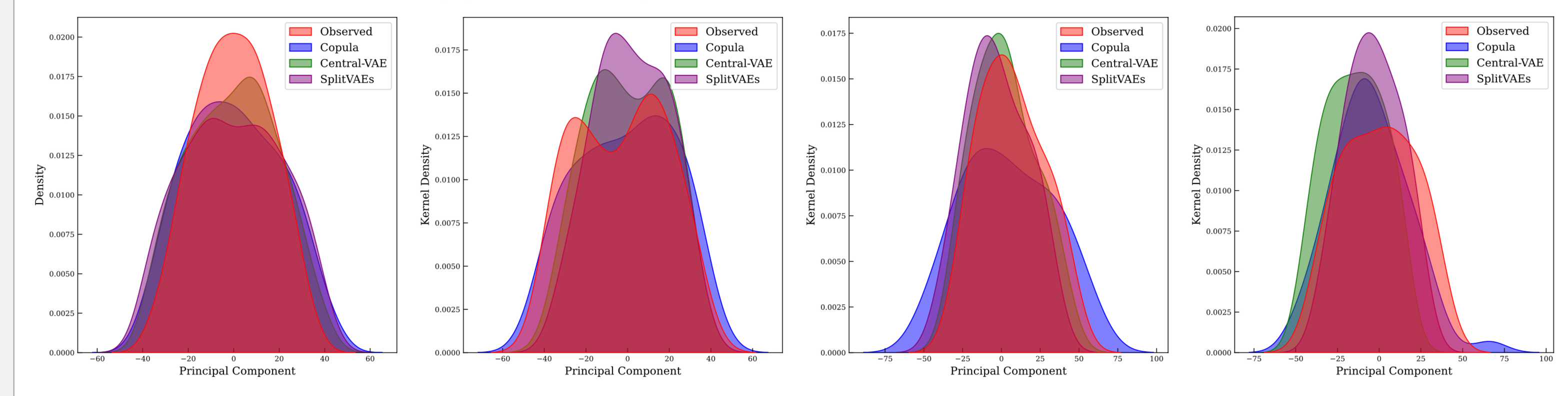
- SplitVAEs demonstrate an **inherent robustness to varying sizes of latent dimensions** in the server VAE and the dimensions of the edge encoders and decoders.

Reduction in Data Movement

- SplitVAEs highlight the **ability of edge devices to generate compact low dimensional embeddings** which is an essential consideration for transmission speed in data-rich environments.
- SplitVAEs **enable reduction in data transmission costs and help save on bandwidth and latency associated** with streaming.

RESULTS

Applicability to Diverse Case Studies

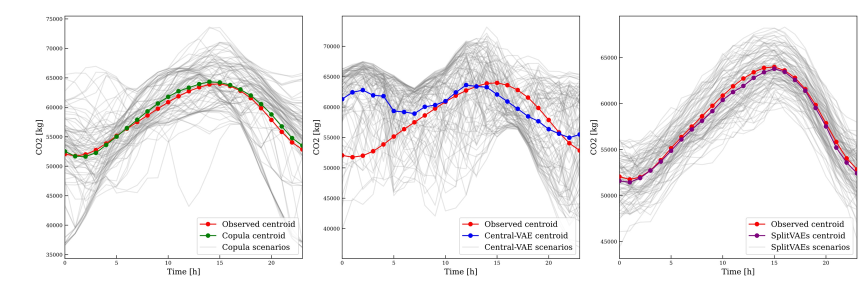


(a) USAID (b) ACES (c) ACTIVSg Demand (d) ACTIVSg Renewable

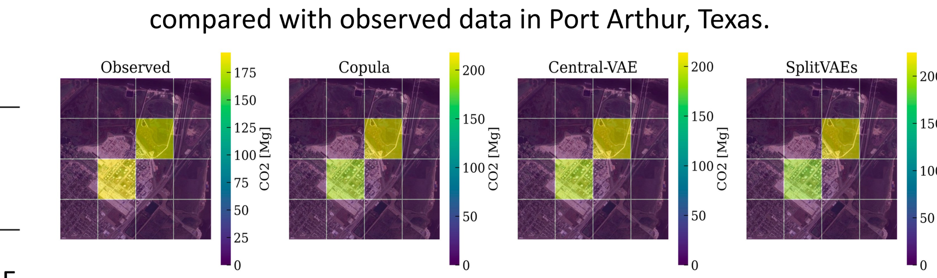
Embedding distributions between observed data and generated scenarios by different methods across four case studies

Dataset	Metric	Copula	Central-VAE	Split-VAE
USAID	FID	2.1768	2.7144	2.7138
	ES	0.0923	0.1570	0.1562
	RMSE	0.2408	0.2374	0.2371
	CRPS	0.2882	0.1144	0.1135
ACES	FID	2.2895	1.9055	0.9713
	ES	0.1129	0.1458	0.1236
	RMSE	0.3036	0.2182	0.1991
	CRPS	0.3094	0.2327	0.1649
Demand	FID	1.5397	1.3039	1.2947
	ES	0.1288	0.1433	0.1287
	RMSE	0.3279	0.2462	0.2419
	CRPS	0.2980	0.1842	0.1824
Renewable	FID	2.1049	1.6767	1.6509
	ES	0.3976	0.3094	0.3066
	RMSE	0.4809	0.3343	0.3351
	CRPS	0.3196	0.1224	0.1219

Benchmarking evaluation metrics (FID, ES,RMSE, CRPS) across four different datasets.

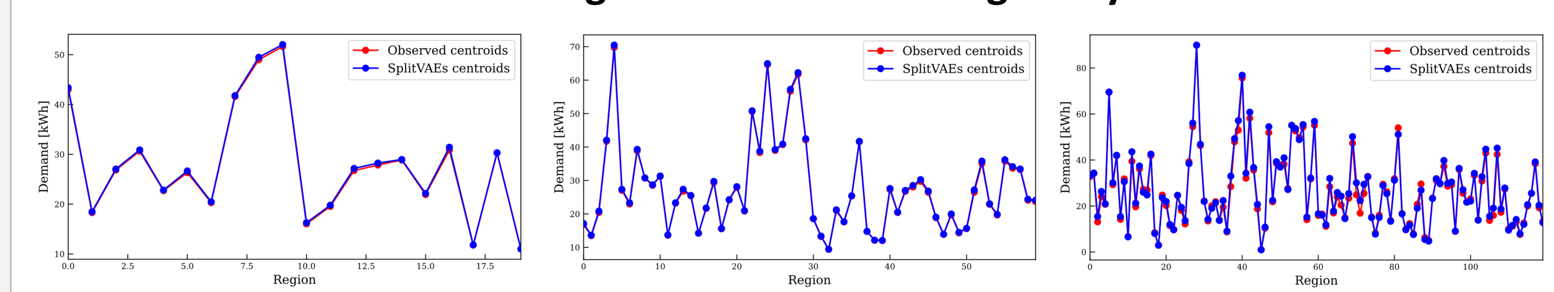


Time series analysis of carbon dioxide emissions scenarios compared with observed data in Port Arthur, Texas.



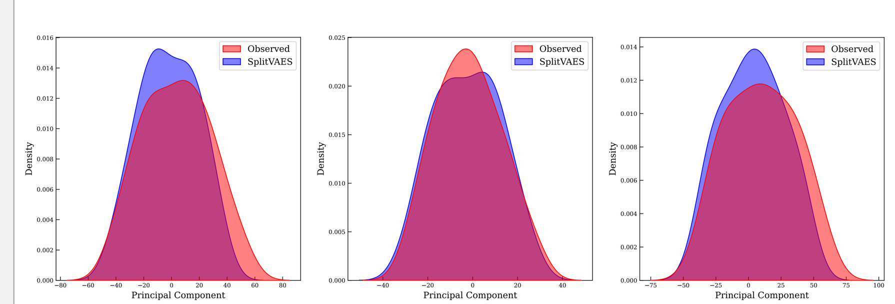
Geographical heatmap illustrating the concentration of carbon emission volume profile in Port Arthur, Texas

Handling Dimensional Heterogeneity



(a) 20 Regions (b) 60 Regions (c) 120 Regions

Time series analysis of scenarios generated in three different region decompositions (20-60-120) compared with observed data for ACTIVSg-Load dataset



(a) 20 Regions (b) 60 Regions (c) 120 Regions

Embedding distributions and metrics between observed data and generated scenarios by different region decompositions

Metric	20-Regions	60-Regions	120-Regions
FID	2.6895	2.7925	2.5433
ES	0.1532	0.1523	0.1825
RMSE	0.2226	0.2238	0.2553
CRPS	0.2009	0.2075	0.2057

Effect of local sub-problem sizes on quality of generated scenarios

Reduction in Data Movement

Dataset	Original Size (MB)	Size of Transmitted Data (MB)			
		20	16	10	8
USAID	68	51	43	38	27
ACES	998	282	201	159	92
Demand	1830	352	281	192	118
Renewable	548	242	168	148	85

Size of data transferred with varying edge-level latent dimension

Metric	20	16	10	8
FID	0.7230	1.7947	3.044	3.6610
ES	0.1489	0.1587	0.1870	0.1888
RMSE	0.2143	0.2419	0.2671	0.2677
CRPS	0.2130	0.2242	0.2359	0.2448

Effect of varying latent dimensions of edge device level

REFERENCES

1. Y. Li, J. Li, and Y. Wang, “Privacy-preserving spatiotemporal scenario generation of renewable energies: A federated deep generative learning approach,” IEEE Transactions on Industrial Informatics, vol. 18, no. 4, pp. 2310–2320, 2021.
2. C. Goncalves, R. J. Bessa, and P. Pinson, “Privacy-preserving distributed learning for renewable energy forecasting,” IEEE Transactions on Sustainable Energy, vol. 12, no. 3, pp. 1777–1787, 2021.
3. J. Liang and W. Tang, “Sequence generative adversarial networks for wind power scenario generation,” IEEE Journal on Selected Areas in Communications, vol. 38, no. 1, pp. 110–118, 2019.
4. M. Jia, C. Shen, and Z. Wang, “A distributed probabilistic modeling algorithm for the aggregated power forecast error of multiple newly built wind farms,” IEEE Transactions on Sustainable Energy, vol. 10, no. 4, pp. 1857–1866, 2018.
5. P. Pinson, H. Madsen, H. A. Nielsen, G. Papaefthymiou, and B. Klockl, ¨ “From probabilistic forecasts to statistical scenarios of short-term wind power production,” Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology, vol. 12, no. 1, pp. 51–62, 2009.

CONCLUSION

- We demonstrate the **applicability and feasibility** of SplitVAEs specifically **in the context of multi-stakeholder infrastructure systems** by leveraging real world datasets.
- We establish the **feasibility and superior computational performance** of our proposed method in various aspects including performance comparison with other approaches, architecture scalability and ease of data movement according the above experiments.
- Overall, SplitVAEs are **a compelling alternative that enables scenario generation** for SO problems in multi- stakeholder-led infrastructure systems without the need to transfer underlying datasets.

1. This project is featured as a finalist for Best Student Paper Competition in IISE 2024 Annual Meetings in Montreal, Canada.
2. This project is under review for The International Conference for High Performance Computing, Networking, Storage, and Analysis (SC2024) in Atlanta, GA.