A comprehensive investigation of variational auto-encoders for population synthesis

Abdoul Razac SANÉ, PhD Student

Laboratoire SPLOTT

Université Gustave Eiffel

22/01/2024

Definition : Synthetic population is a generic representation of a group of individuals or households, that mimics the characteristics and behaviors of the actual population.

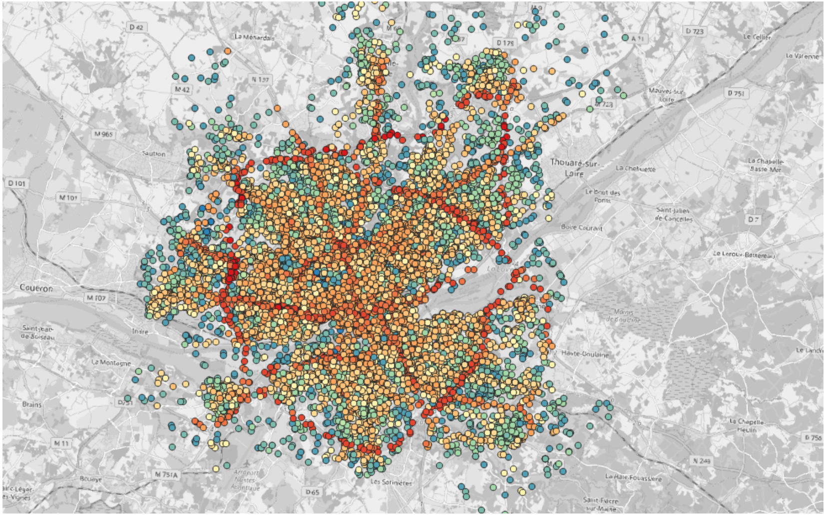
⇒ **Activity-based Multi-Agent Systems (MAS)**

- Fine-tuned assessment
- Space-time tracking
- Focus on specific classes

→ Many fields, such as social science research, urban planning, public health and transportation modeling

The approach increasingly used in urban logistic modeling

- Sakai et al. [2020] : SimMobility Freight : An agent-based urban freight simulator for evaluating logistics solutions
- De Bok et al. [2022] : Application of an empirical multi-agent model for urban goods transport to analyze impacts of zero emission zones in The Netherlands
- Belfadel et al. [2023] : A conceptual digital twin framework for city logistics


Couëron
Thouaré-sur-Loire
Haute-Goulaine
Les Sorinières
Saint-Jean-de-Boiseau
La Montagne
Brains
Bouaye
Sautron
Indre
La Chebuette
Boire Courant
Le Loroux-Bottereau
La Chapelle-Heulin
La Haie-Fouassière
Saint-Fiacre-sur-Maine
Mauves-sur-Loire
Le Cellier
La Varenne
D 42
D 101
M 101
M 107
M 965
M 42
N 137
D 178
D 723
D 751
D 7
D 65
M 11
M 751A
N 249
La Ménardais

- Representative population data is required for these models
- In practice, there is generally little data available to study certain problems.

Quality of simulator inputs $\Rightarrow$ Modeling accuracy

Problem :

How can we generate a synthetic population that best represents the real population ?

Recently, new methods for generating synthetic populations based on deep learning, notably Variational Autoencoders (VAE), have been developed. Such methods serve to overcome the limitations of traditional methods, such as IPF or hierarchical models.

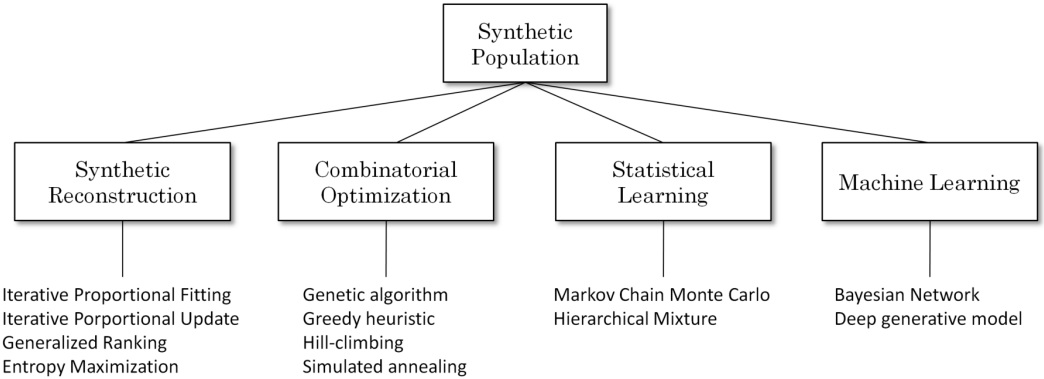
The aim of this work is to propose an alternative to the traditional population synthesis method by providing the practitioner with a practical guide to generating a synthetic population using a VAE.

More specifically, the aim is to provide :

- An accessible theoretical explanation of how VAEs function.
- An evaluation of training sample sizes
- An evaluation indicators necessary to guarantee high-quality results

Plan

Population synthesis methods



The Synthetic Reconstruction and Combinatorial Optimization methods yield synthetic populations by means of replicating individuals

→ **Synthetic Reconstruction (SR)** : deterministic method

→ **Combinatorial Optimization** : stochastic method

- **Advantages** : Easy to implement and ensure that aggregate synthetic population values match marginal data
- **Limitations** : it requires a large number of individuals and it encounters difficulties when taking into account a large number of attributes, difficulty in handling continuous, Sampling-zero problem

Statistical and Machine Learning methods generate a population by following a joint probability estimation

→ Statistical Learning

- **Advantages** : it incorporates both causal relationships and probabilistic semantics
- **Limitations** : the curse of dimensionality, the need for a large sample size, the assumption of conditional independence, and the difficulty in handling continuous, Sampling-zero problem

→ Machine Learning

- **Advantages** : can handle high dimensional data, able to learn the joint distribution of numerical and categorical attributes, takes into account sampling zero problem
- **Limitations** : Computing time

Generative adversarial network (GAN)

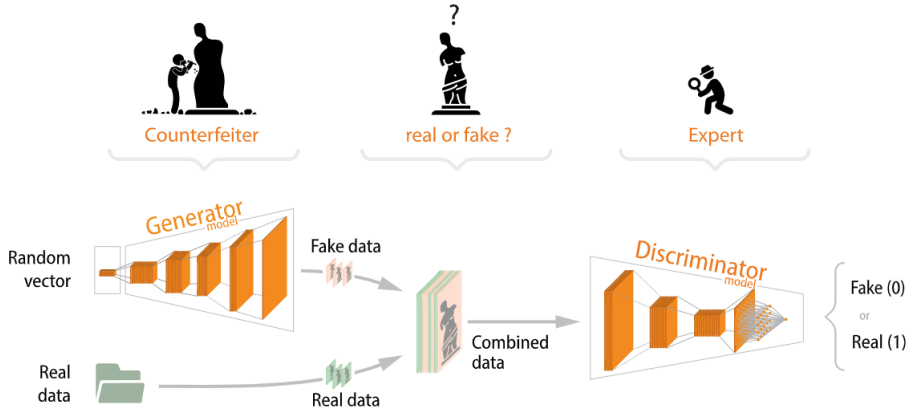


Figure – GAN illustration, Fidle(2023)

Variational AutoEncoder (VAE)

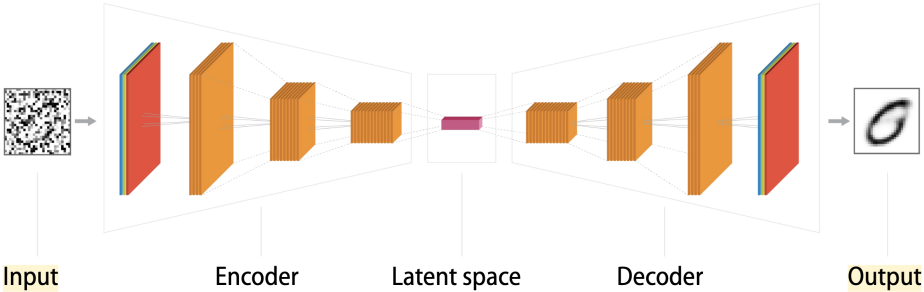


Figure – VAE illustration, Fidle(2023)

Figure – The autoencoder is an unsupervised model trained to copy its inputs. It compresses the input vector X with dimensions n into the latent representation Z with fewer dimensions and then reconstructs it back into the data space

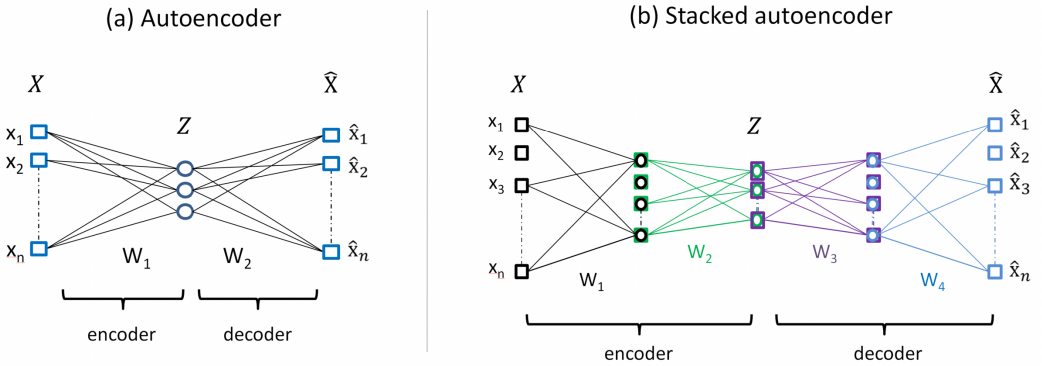
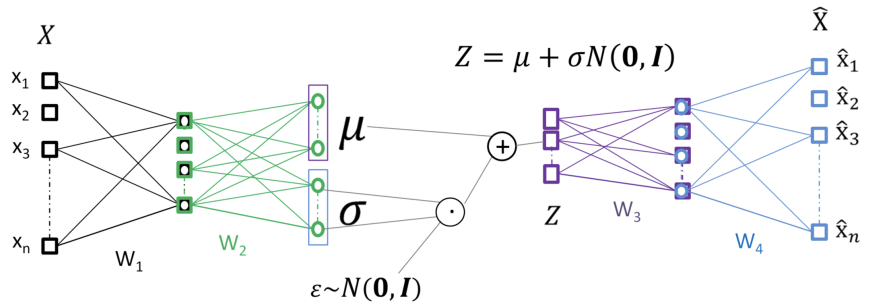


Figure – VAE illustration and loss function



$$L(\phi,\theta) = \sum_{i=1}^{N} \underbrace{c_1 \times \sum_{k\in\{num\}} (\hat{x}_i^k - x_i^k)^2}_{\text{Numerical cost}} + \underbrace{c_2 \times \sum_{k\in\{cat\}} \sum_{d=1}^{D_i} x_i^{kd} \log \hat{x}_i^{kd}}_{\text{Categorical cost}} + \underbrace{\beta \times \mathbb{KL}(\mathcal{N}(\mu^k, \sigma^k) \| \mathcal{N}(\mathbf{0}, \mathbf{I}))}_{\text{KL cost}}$$

→ **To yield Loire-Atlantique pop.**

→ **Around 1.4M**

Nantes HTS data

- 12,700 households
- 29,500 people aged 5 and older

Train and test datasets

- 80% → model-building dataset
- 20% → test dataset

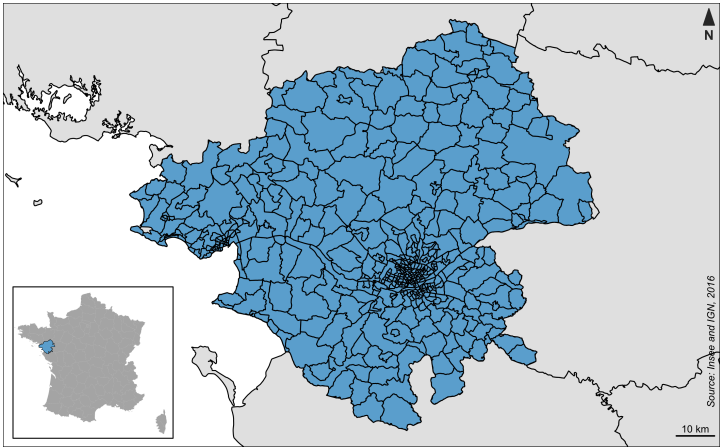


Table – Overview of household and individual features

Groups	Features	Type[1]	N of values[2]	Description
Household (HH)				
	house_type	cat	5	House type
	house_occupation_type	cat	4	House occupation type
	household_size	num	-	Number of persons in the HH
	number_of_vehicles	num	-	Number of vehicules in the HH
	number_of_bikes	num	-	Number of bikes in the HH
	has_internet	cat	2	Has internet access
Individuals				
	link_ref_person	cat	5	Link with the HH reference person
	age	num	-	age
	sex	cat	2	Sex
	is_adolescent	cat	2	Is adolescent
	school_level	cat	5	School level
	employed	cat	2	Is employed
	studies	cat	2	Studies
	socioprofessional_class	cat	9	Socioprofessional class
	has_license	cat	2	Has a driver's license
	number_of_trips	num	-	Number of previous day's trips

[1] Indicates whether the feature is numerical (num) or categorial (cat)

[2] Number of possible values, it is only relevant for categorical features

$\rightarrow$ A set of models by combining different values of hyperparameters : 1,512 **models**

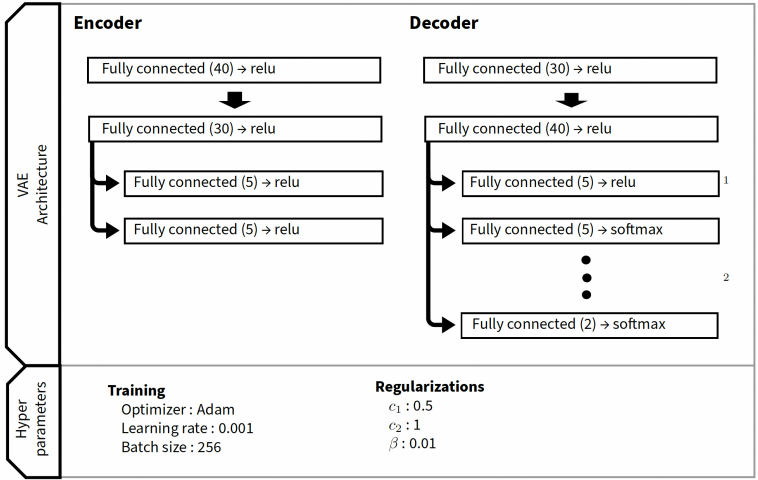
Table – Overview of models hyper-parameters

Number of layers	Number nodes	Latent dim	Mini-Batch size	c_1	c_2	β	Gradient descent algo.	Node activations
1 ; 2 ; 3	60 ; 40 ; 30 ; 20	5 ; 10 ; 15	128 ; 256 ; 512 ; 1,024	0.4 ; 0.5 ; 0.7	1	0.01 ; 0.05 ; 0.005	adam	ReLU ; tanh

$$SRMSE = \frac{\sqrt{\frac{1}{N}\sum_k (\pi_k - \hat{\pi_k})^2}}{\frac{1}{N}\sum \pi_k} \quad (1)$$

$$Corr = \frac{cov(\pi, \hat{\pi})}{sd(\pi) \times sd(\hat{\pi})} \quad (2)$$

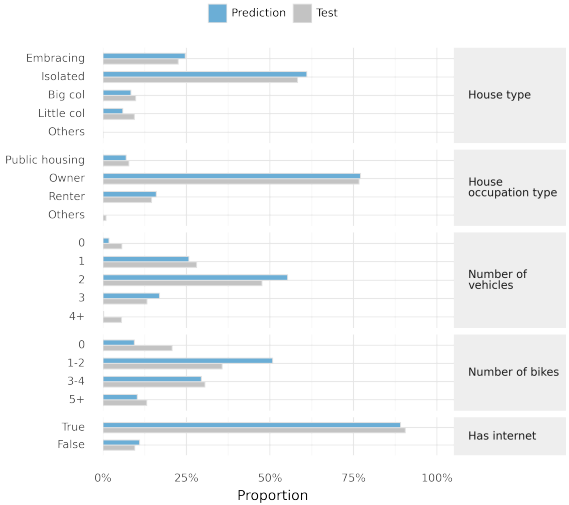
Best model



[1] : Five numerical attributes: one input/output for each
[2] : Other categorical attributes

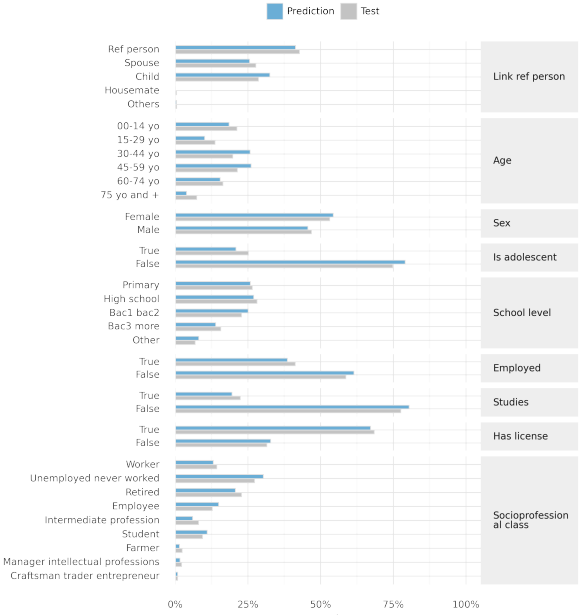
Results : Performances on the household attributes

An error below 3%, with the exception of certain modalities of the variables "number of vehicles" and "number of bicycles"



Results : Performances on the individual attributes

An error of less than 4% for all attributes except Age. The age categories of age between 15-29 years old and the 75+ years old are underestimated with 5% less than the true population, whereas the 30-44 and 45-59 age categories are overestimated by 5%compared to the test set



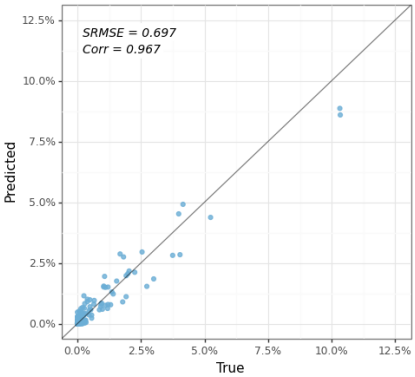


Figure – Household attributes : type of housing, type of house occupancy, number of vehicles and Internet availability

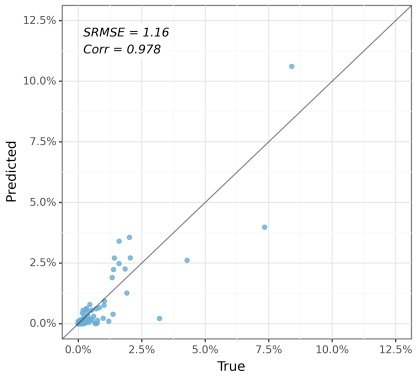
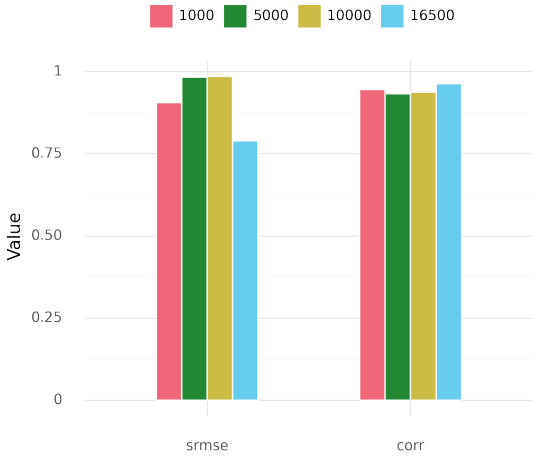


Figure – Individuals attributes : Gender, Educational level, Socio-professional categories and Link with the reference person

Results : Training size influence

Figure – Values of different metrics on the 41.6 k-dimensional joint representation with a growing training set size



Sampling zero :

$\hat{\pi} \in$ Test et $\hat{\pi} \notin$ Train

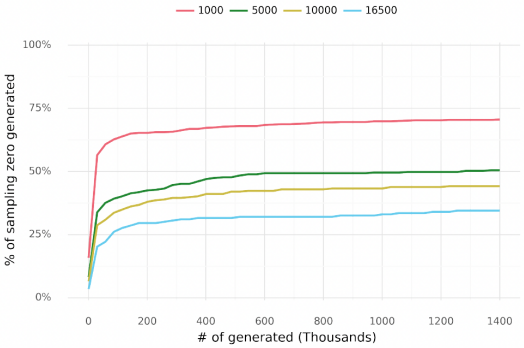
Structural zero :

$\hat{\pi} \notin$ Test et $\hat{\pi} \notin$ Train

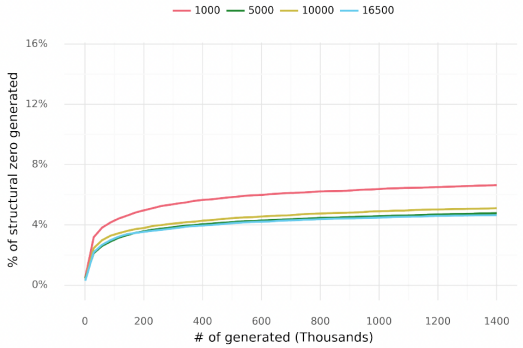
Table – Total number of sampling zeros and structural zeros and Fraction of agents whose cross modalities are sampling zeros and structural zeros over the 1.4M generated agents according to the training size (18k-dimensional joint)(27k-dimensional joint)

Train size	Number of possible cross-modalities		Fraction of agents with these cross-modalities	
	# sampling	# structural	% of agent with sampling	% of agent with structural
1,000	702	25,979	20.96	16.52
5,000	426	25,797	6.48	12.96
10,000	324	25,611	2.55	10.31
16,500	203	25,244	1.37	9.59

Figure – Fraction of total sampling zeros and total structural zeros generated in the synthetic population for each VAE and for a growing total population size

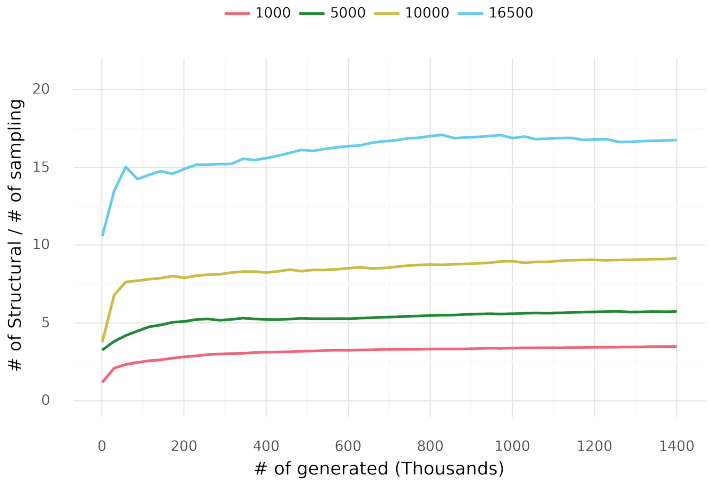


(a)



(b)

Figure – Ratio between the number of structural zeros to sampling zeros versus the size of the generated population



The objectives of this study have been to provide a methodological framework to implement and evaluate a VAE model aimed at generating a synthetic population of agents using mixed data.

☺ Remarkable performance of VAEs

- The average errors for attribute marginal data were less than 3 percentage points
- Ability to generate new individuals not present in the training sample
- Take into account mixed data, but categorical attributes becomes more robust.
- Capable of processing high-dimensional data while maintaining good performance with a reasonable sample size.

Université Gustave Eiffel

Main limitation

- Structural zeros : the use of VAE requires post-processing to eliminate them.

Perspectives ...

- Making the code available to the community
- Taking marginal data into account to improve model performance
- To generate individuals in households,
- Integrating sequential data : activity plans including consumer-side logistics

Thank you for your attention ...

Bibliographie

Takanori Sakai, André Romano Alho, BK Bhavathrathan, Giacomo Dalla Chiara, Raja Gopalakrishnan, Peiyu Jing, Tetsuro Hyodo, Lynette Cheah, and Moshe Ben-Akiva. Simmobility freight : An agent-based urban freight simulator for evaluating logistics solutions. *Transportation Research Part E : Logistics and Transportation Review*, 141 : 102017, 2020.

Michiel De Bok, Lóránt Tavasszy, and Sebastiaan Thoen. Application of an empirical multi-agent model for urban goods transport to analyze impacts of zero emission zones in the netherlands. *Transport Policy*, 124 :119–127, 2022.

Abdelhadi Belfadel, Sebastian Hörl, Rodrigo Javier Tapia, Dimitra Politaki, Ibad Kureshi, Lorant Tavasszy, and Jakob Puchinger. A conceptual digital twin framework for city logistics. *Computers, Environment and Urban Systems*, 103 :101989, 2023.