Developing digital twins of infrastructure for risk analysis

Aaron Dunton
PhD Candidate, Dept. of Civil Engineering, University of Illinois, Urbana-Champaign, United States

Paolo Gardoni
Alfredo H. Ang Family Professor, Dept. of Civil Engineering, University of Illinois, Urbana-Champaign, United States

ABSTRACT: Infrastructure are vulnerable to natural and anthropogenic hazards. The general steps in the risk analysis of infrastructure include modeling the undamaged network, modeling the hazard, predicting the damage to components, and assessing the functionality of the network given the predicted damage. However, these steps require information about the infrastructure that might not be available. This work develops a general procedure to generate representative and functionally equivalent models (i.e., digital twins) of infrastructure. The proposed procedure is systematic and uses detailed geospatial information that are generally available. By using only these data, representative networks can be generated for most locations of interest. We also develop a general modeling approach to assess the impact of localized damage on infrastructure networks. We consider the uncertainty in the network topology by generating multiple representative network realizations. Finally, we demonstrate the proposed procedures by generating representative wastewater networks and assessing the impact of localized damage on the network for a case study location.

Critical infrastructure transport essential resources to buildings. In the case of hazardous events, damage to critical infrastructure may cause disruption. Risk analysis requires predictions about the impact of future hazardous events on the infrastructure. To make these predictions, models are used. Specifically, networks are models that describe the connection of components and how the components act as a system. However, the information needed to define the network (i.e., the location of components and how they are connected) may not be known. For example, network information might not be available when the infrastructure is controlled by a different government or corporation. However, the impact of hazardous events on unknown infrastructure may still be of interest. For example, a corporation may wish to model the impact of future hazardous events on the operations of another corporation on which their operations are dependent. For modeling unknown infrastructure, network generators use other data that are available to create representative networks.

Existing infrastructure network generators have not focused on data availability. Some studies have not used data that is generally available. In these studies, synthetic/virtual infrastructure networks have been developed as research tools, either as test beds or for anonymizing sensitive network data. These include generators for elector power networks (Aksoy et al. 2019, Birchfield et al. 2017), sewer networks (Urich et al. 2010, Jeffers and Montalto 2018), potable water networks (Möderl et al. 2011), natural gas networks (Vaccariello et al. 2020), and for multiple types of networks (Sitzenfrei et al. 2010, Wang et al. 2022). These studies do not consider detailed location-specific data that could provide additional constraint to generate more realistic networks for location-specific risk analyses. Other studies have used data that is not generally available. The location of demand nodes is often assumed (Ahmad et al. 2020, Chahinian et al. 2019, Sitzenfrei et al. 2020). Other studies generate cost-optimal networks (Wang et al. 2017, Moeini and Afshar 2019, Chahinian et al. 2019, Sitzenfrei et al. 2020), but the detailed cost data needed for these models is not generally available. This data is also location and time-specific. Moreover, optimization may not represent typical design procedures. In general, there is a need for a location-specific infrastructure network generator that uses data that should always be available. This network generator would provide a baseline representation that could always be used.

In this paper, we use generally available data to generate models of infrastructure for risk analysis. Critical infrastructure typically connect centralized production, storage, or collection facilities with every building in a community. This produces a hierarchy in the network, whereby larger components serving larger areas are different than the infrastructure that connect to individual buildings. We call the larger components that serve large areas of a community the arterial infrastructure, and we call the smaller components that connect to every building the capillary infrastructure. For community-level risk analysis, the functionality of arterial infrastructure is more important than that of the capillary infrastructure. In this paper, we propose a procedure to generate arterial infrastructure networks using generally available data. We use this procedure for risk analysis of unknown infrastructure to localized damage. That is, we model the functionality of the infrastructure when we know where damage will occur, but we do not know anything about the infrastructure. In this situation, uncertainty in the network itself must be considered. We modify the proposed arterial network generator to account for this uncertainty, generating many representative network realizations. We demonstrate the proposed procedure by generating arterial networks and assessing the impact of localized damage on the wastewater network of a medium-sized city in the United States.

The rest of this manuscript is organized as follows. In Section 1, we introduce the generally available data that we use to generate arterial infrastructure. In Section 2, we present the proposed procedure. In Section 3, we describe a procedure to consider uncertainty in the network generator. Finally, in Section 4, we present the results for the case study location.

1. GENERALLY AVAILABLE DATA

In this section, we describe the data that we use to generate arterial networks. The component locations and connections between components are in general not available. However, there are many detailed geospatial data that are available for most locations. These data typically come from remote sensing or from large, standardized data collection efforts. By using only these generally available data, we are able to generate representations of the infrastructure for most locations of interest without any direct information on the infrastructure. Table 1 lists these data and the sources that we use. These datasets have consistent coverage for the entire United States, and comparable datasets should be available for international locations. To manipulate the geospatial inputs, we use GeoPandas (Jordahl et al. 2021).

Table 1: Inputs for the arterial network generator

Input	Description	Source
Streets shapefile	Mainline streets are potential locations for arterial infrastructure	Boeing (2017)
Building footprints shapefile	Locations where arterial networks provide resources	Microsoft (2018)
Digital elevation model	Used to evaluate weights for determining the topology	USGS (2022)
Total area shapefile	Area over which the network is being generated	Manually specified
Primary node	Location from which resources are being distributed or to which waste is being collected	Manually specified

Before proceeding with the proposed procedure, these data are cleaned and pre-processed. First, we manually modify the street shapefile so that there is topological consistency. For example, street intersections are represented by two lines with endpoints that are exactly coincident. Occasionally, the two endpoints in the unmodified shapefile are not at the exact same location. In this case, we modify the endpoints to be exactly coincident. Second, we merge any boulevards in the street shapefile into single lines. That is, the street shapefile may have two parallel lines representing a single street along which there is likely only one infrastructure line.

The rest of the pre-processing of the inputs more precisely defines the extent of the network that is generated. First, we remove any buildings without associated streets or streets without associated buildings. For example, if the capture date of the building data is after the capture date of the street data, it is possible there are buildings in locations that are served by streets that were not present when the streets were recorded. More accurate networks could be developed by manually adding missing streets and buildings using more recent aerial imagery. However, we suggest that removing these parts of the network produces a reasonable representation assuming all data is the most recent available, and any new additions to the network are relatively small as compared to the rest of the network. Second, we remove any small or geographically isolated buildings that would not be connected to the infrastructure. Third, we remove any lines from the street shapefile that represent interstates or large highways along which infrastructure would not be present. Finally, we create the total area shapefile to represent the likely geographic extent of the network based on the other inputs.

2. PROPOSED ARTERIAL NETWORK GENERATOR

The proposed procedure has three steps that we illustrate in Figure 1. First, we break the full area into subareas. The arterial network connects the subareas. Second, we determine the point where the infrastructure in each subarea connect to the arterial infrastructure. Finally, we determine the topology of the arterial infrastructure by connecting all of the subareas to the primary node. The rest of this section is organized by these steps.

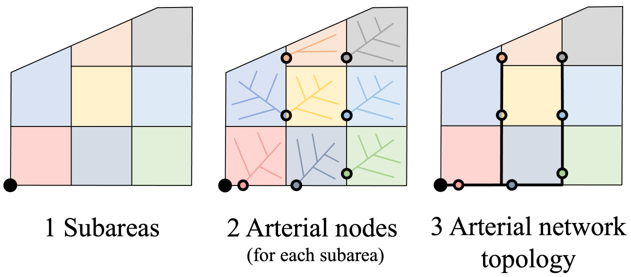

Figure 1: Steps of the proposed procedure to generate arterial networks

2.1 Step 1: Identify subareas

We identify preliminary subareas based on mainlines in the street network. Any large streets that run continuously through large portions of the network area are manually identified as street mainlines. These street mainlines can usually be easily identified. We assume that all arterial infrastructure is along the identified street mainlines and there is no capillary infrastructure along the street mainlines. So, each street mainline is a boundary for the adjacent subareas (i.e., no subarea can cross a street mainline). In this way, the street mainlines define preliminary subareas. The preliminary subareas may be larger or smaller depending on the layout of the mainline street network. However, ideally the subareas should all be a similar size. That is, instead of ensuring that the subareas are all equally sized, we propose to use the street mainlines that we can easily identify to realistically divide the total area into subareas. Consequently, the subareas have no clear physical meaning in the proposed procedure.

The preliminary subareas are then further refined. In each subarea, we remove lines that connect pieces of the network that should be disconnected. Also, we add non-mainline extensions to isolated groups of buildings that would be directly connected to a mainline. For example, this is common for apartment complexes and for large commercial and industrial building.

In these cases, the pathways along which the infrastructure lines would be placed are commonly not streets and are not included in the available street shapefiles. Given these modifications, the street network in each subarea may be disconnected. We define one final subarea for each connected components of the street network in each preliminary subarea.

2.2 Step 2: Identify arterial nodes

Next, we determine the location where the infrastructure in each subarea are connected to the arterial infrastructure. To do this, we use a graph. A graph is a pair of a vertex/node set and an edge set, $G = (V, E)$, where $E \subseteq V^2$. An edge (i, j) of a graph represents a connection between node i and node j. Each edge has a weight that represents how difficult it is to move between the nodes. A natural choice for the weight when the edges are representing street segments is distance. More specific weights can be formulated for specific circumstances. For example, for wastewater networks, downhill flow is much easier than uphill flow. So, the elevation can be considered in the formulation of the weights. We propose edge weights for generating wastewater networks as follows:

$$w_{ij} = \frac{h_j - \min(h)}{\max(h) - \min(h)} d_{ij} \tag{1}$$

where h_j is the elevation of node j, $\min(h)$ is the minimum of all nodal elevations, $\max(h)$ is the maximum of all nodal elevations, and d_{ij} is the distance between node i and node j. With this formulation, more weight is given to edges that go to higher elevation. So, flow in the downhill direction is encouraged.

The mainline street network can be modeled as a graph $S = (V_s, E_s)$ where V_s are street intersections and E_s are street segments that connect street intersections. For the arterial network, there are additional nodes that connect to each of the subareas. Adding these nodes to S yields the modified mainline street network, $S' = (V_s', E_s')$, where $V_s \subseteq V_s'$. To determine the arterial node for each subarea, we choose from the points where the non-mainline streets in the subarea intersect the mainline streets. We choose the candidate that is closest to the primary node in S, considering the weight of the edges. We use NetworkX for this, and other, graph operations (Hagberg et al. 2008). Finally, $E_s{}'$ is determined. The topology from S is maintained, but each edge in E_s may be split if a subarea node is now in between the two street intersections.

2.3 Step 3: Identify arterial network topology

Finally, we determine the arterial network, $N = (V_n, E_n)$. We propose to model the arterial infrastructure as a subtree of S' that spans all of the subarea nodes. So, $V_n \subseteq V_s{}'$ and $E_n \subseteq E_s{}'$. For each subarea node, we determine the shortest path to the primary node in S', considering the edge weights. The nodes and edges along the path are added to V_n and E_n, respectively. So, all subarea nodes are included in N. But, if a street intersection node is not used to connect one of the subarea nodes to the primary node, then it is not included in N.

3. NETWORK UNCERTAINTY FOR LOCALIZED DAMAGE

In this section, we consider the use of the arterial network generator for assessing the impact of localized damage on the network. For example, this type of damage could come from accidents, attacks, or earthquake-induced liquefaction. Whatever the source, we consider the case where the geographic extent of damage is known. If we also know the location of the infrastructure, then we know which components are damaged (i.e., the components in the damage area). If we do not know anything specific about the location of the infrastructure components, we use the proposed arterial network generator. However, in this case we need to consider uncertainty in the network itself. We do this by generating many realizations of the network topology, step 3 in the proposed procedure. We suggest that the topology is the most important step in terms of the connectivity of each subarea.

We illustrate this in Figure 2. Regardless of the exact geographic extent of each subarea and where each subarea is connected to the arterial network (the other steps in the proposed procedure), there is some arterial path from the subarea to the primary node. Whether components along this path are damaged determines whether the area is disconnected. There is uncertainty because subareas can be connected to the primary node with multiple paths of similar length but that pass through different locations. Strictly, this uncertainty should always be considered, though it may not be important for some hazards that do not have sharp variation in intensity. This comes from the assumptions we made in the deterministic procedure described in Section 2. That is, the actual designer of the infrastructure does not use a deterministic procedure like the one that we proposed, though the formulation accounts for the correct design tendency.

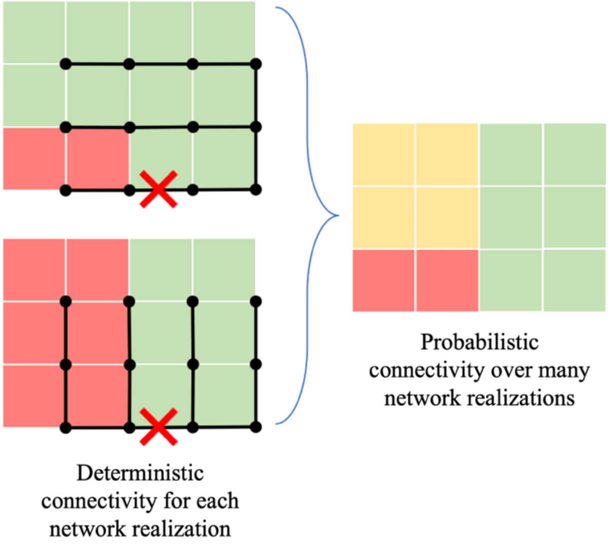
Probabilistic connectivity over many network realizations
Deterministic connectivity for each network realization

Figure 2: Uncertainty in the arterial topology and how it affects the analysis results

We consider this uncertainty by modifying the weights used in the shortest path tree algorithm. We apply multiplicative random error to the weights as follows:

$$w_{ij}' = \delta w_{ij} \tag{2}$$

To determine the value of δ for each edge, we generate independent samples from a lognormal distribution. We consider two cases for the parameters of the lognormal distribution. First, we consider a low uncertainty (LU) case with a mean of 1 and a standard deviation of 0.05. Second, we consider a high uncertainty (HU) case with a mean of 1 and a standard deviation of 0.25.

4. CASE STUDY

Following the proposed procedure to generate arterial infrastructure networks, we generate wastewater networks for Irving, Texas. The location of Irving is shown in Figure 3. Irving is a city with approximately 250,000 residents. We collect and pre-process the input data, shown in Figure 4. Figure 4 also shows the damage area for which we are assessing the risk to the wastewater network.

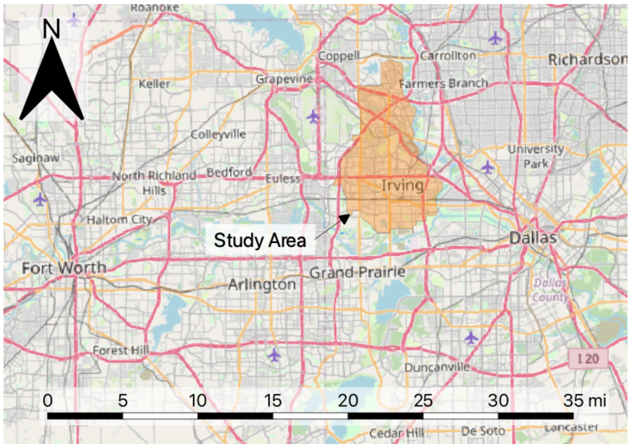

Figure 3: Case study location: Irving, Texas

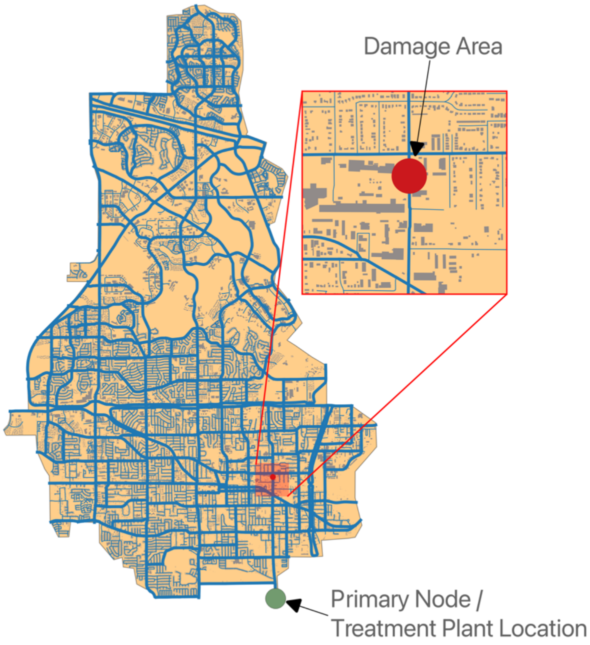

Figure 4: Input data for Irving, Texas

We generate arterial wastewater networks for this location as described in Section 2. We use the weight formulation from Equation 1. To assess the impact of localized damage on the functionality of the wastewater network, we use the procedure described in Section 3. Figures 5 and 6 map the probability of disconnection of each subarea for the LU and HU cases, respectively. We evaluate the probability of disconnection as the number of network realizations where the subarea is disconnected divided by the total number of network realizations. We generate enough realizations of the network so that the mean coefficient of variation of the estimates of the probability of disconnection, over the subareas with non-zero probability, is less than 0.5. Figures 5 and 6 demonstrate that the extent of disconnection may be larger for the HU case. However, some subareas have lower probability of disconnection for the HU case. These subareas are further away from the damage area, so with more uncertainty there is more chance that these areas may be connected to the primary node with an alternative path that is not damaged. Overall, the two maps show similar subareas that will likely be disconnected.

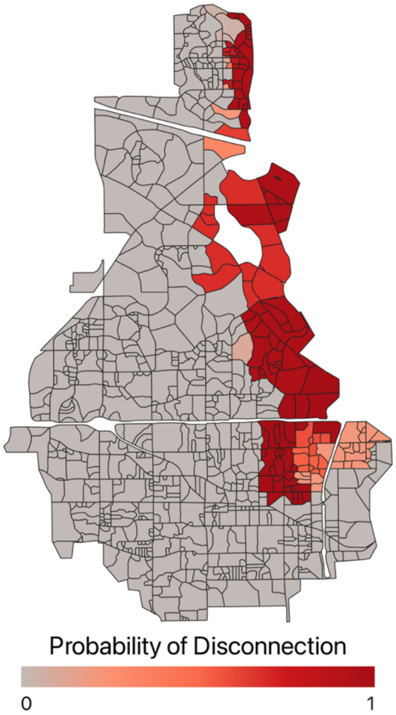

Figure 5: Probability of disconnection of each subarea for the LU case

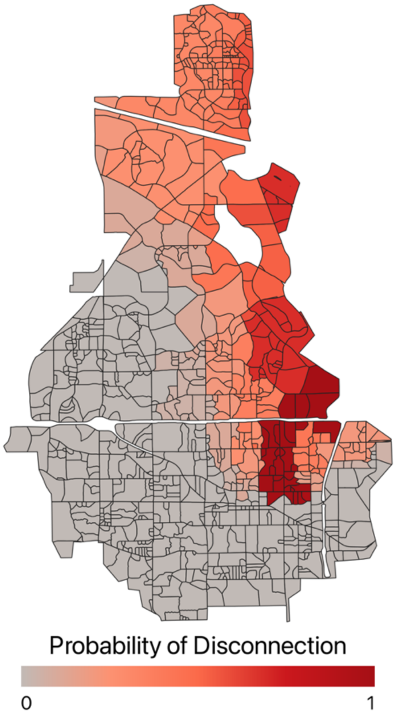

Figure 6: Probability of disconnection of each subarea for the HU case

Figures 7 and 8 plot the histogram of the fraction of the total network area that is disconnected for each realization of the network. These figures aggregate the results for each network realization into a single quantity and show how that quantity varies over the network realizations. The HU case has slightly more damage on average. Also, the variation is much higher for the HU case.

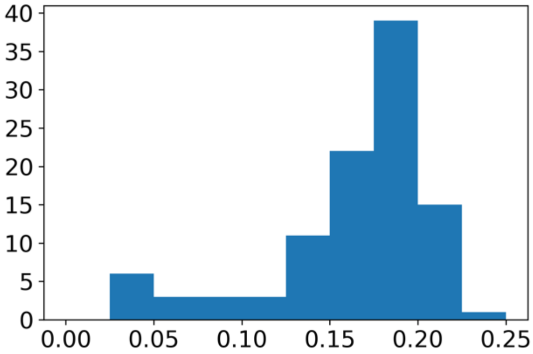

Figure 7: Histogram of the percentage of the network area that is disconnected for each of the network realizations, in the LU case

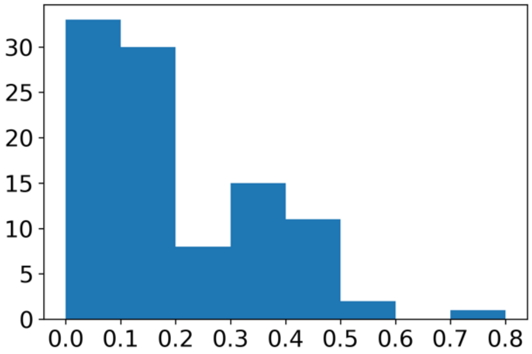

Figure 8: Histogram of the percentage of the network area that is disconnected for each of the network realizations, in the HU case

5. CONCLUSION

We proposed a systematic procedure to generate representative models of infrastructure using generally available data. In particular, we defined arterial infrastructure as the larger components of the infrastructure that serve large areas in a community. To generate the arterial infrastructure, we proposed a three-step procedure. First, we divided the total area into subareas. Second, we determined nodes where each subarea connects with the arterial infrastructure. Third, we determined the network topology, connecting the subarea nodes with the primary node. Then, we used the network generator to assess the impact of localized damage on an unknown network. We generated multiple network realizations and evaluated the probability of disconnection over the realizations. Finally, we applied the proposed procedure to a case study location. We tested low uncertainty and high uncertainty cases when generating multiple network realizations. Ultimately, both analyses indicate similar subareas that would probably be disconnected given the geographic extent of damage that we started with.

6. REFERENCES

Ahmad, N. Chester, M., Bondank, E., Arabi, M., Johnson, N., and Ruddell, B. L. (2020). A synthetic water distribution network model for urban resilience. *Sustainable and Resilient Infrastructure.* https://doi.org/10.1080/23789689.2020.1788230.

Aksoy, S. G., Purvine, E., Cotilla-Sanchez, E., and Halappanavar, M. (2019). A generative graph model for electrical infrastructure networks. *Journal of Complex Networks* 7(1):128-162.

Birchfield, A. B., Gegner, K. M., Xu, T., Shetye, K. S., and Overbye, T. J. (2017). Statistical considerations in the creation of realistic synthetic power grids for geomagnetic disturbance studies. *IEEE Transactions on Power Systems* 32(2):1502-1510.

Boeing, G. (2017). U.S. Street Network Shapefiles, Node/Edge Lists, and GraphML Files. Harvard Dataverse, ver. 2. https://doi.org/10.7910/DVN/CUWWYJ.

Chahinian, N., Delenne, C., Commandré, B., Derras, M., Deruelle, L., and Bailly, J. (2019). Automatic mapping of urban wastewater networks based on manhole cover locations. *Computers, Environment, and Urban Systems* 78:101370.

Hagberg, A. A., Shult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux, G., Vaught, T., and Millman, J. (eds) *Proceedings of the 7th Python in Science Conference,* SciPy 2008.

Jeffers, S. M., and Montalto, F. (2018). Modeling urban sewers with artificial fractal geometries. *Journal of Water Management Modeling* 26:C445.

Jordahl, K., et al. (2021). Geopandas/geopandas: v0.10.2. doi:10.5281/zenodo.5573592.

Microsoft (2018). U.S. Building Footprints. Microsoft Maps, ver. 2. https://github.com/microsoft/USBuildingFootprints. (May 1, 2022).

Möderl, M., Sitzenfrei, R., Fetz, T., Fleischhacker, E., and Rauch, W. (2011). Systematic generation of virtual networks for water supply. *Water Resources Research* 47(2):W02502.

Moeini, R., and Afshar, M. H. (2019). Extension of the hybrid ant colony optimization algorithm for layout and size optimization of sewer networks. *Journal of Environmental Informatics* 33(2):68-81.

Sitzenfrei, R., Fach, S., Kinzel, H., and Rauch, W. (2010). A multi-layer cellular automata approach for algorithmic generation of virtual case studies: VIBe. *Water Science and Technology* 61(1):37-45.

Sitzenfrei, R., Wang, Q., Kapelan, Z., and Savić, D. (2020). Using complex network analysis for optimization of water distribution networks. *Water Resources Research* 56:e2020WR027929.

Urich, C., Sitzenfrei, R., Möderl, M., and Rauch, W. (2010). An agent-based approach for generating virtual sewer systems. *Water Science and Technology* 62(5):1090-1097.

USGS (2022). Digital Elevation Model. U.S. Geological Survery (USGS): 3D Elevation Program. https://apps.nationalmap.gov/downloader/#/. (May 1, 2022).

Vaccariello, E., Leone, P., and Stievano, I. S. (2020). Generation of synthetic models of gas distribution networks with spatial and multi-level features. *Electrical Power and Energy Systems* 117:105656.

Wang, Q., Savić, D, and Kapelan, Z. (2017). GALAXY: A new hybrid MOEA for the optimal design of water distribution systems. *Water Resources Research* 53:1997-2015.

Wang, Y., Yu, J., Baroud, H. (2022). Generating synthetic systems of interdependent infrastructure networks. *IEEE Systems Journal* 16(2):3191-3202.