Incomplete Data and Technological Progress in Energy Storage Technologies

Sertaç Oruç[1], Scott W. Cunningham[1], Christopher Davis[2], Bert van Dorp[1]

[1] *s.oruc@tudelft.nl, s.cunningham@tudelft.nl, bertvandorp@gmail.com*
[1]Delft University of Technology, Faculty of Technology, Policy and Management, Jaffalaan 5 C2.010, 2628 BX Delft (The Netherlands)

[2] *c.b.davis@rug.nl*
University of Groningen, Center for Energy and Environmental Sciences (IVEM), Nijenborgh 4, 9747AG Groningen (The Netherlands)

Abstract

Energy storage is an important topic as many countries are seeking to increase the amount of electricity generation from renewable sources. An open and accessible online database on energy storage technologies was created, incorporating a total of 18 energy storage technologies and 134 technology pages with a total of over 1,800 properties. In this database information on technical maturity, technology readiness level and forecasting is included for a number of technologies. However, since the data depends on various sources, it is far from complete and fairly unstructured. The chief challenge in managing unstructured data is understanding similarities between technologies. This in turn requires techniques for analyzing local structures in high dimensional data. This paper approaches the problem through the use and extension of t-stochastic neighborhood embedding (t-SNE). t-SNE embeds data that originally lies in a high dimensional space in a lower dimensional space, while preserving characteristic properties. In this paper, the authors extend the t-SNE technique with an expectation-maximization method to manage incompleteness in the data. Furthermore, the authors identify some technology frontiers and demonstrate and discuss design trade-offs and design voids in the progress of energy storage technologies.

Conference Topic

Mapping and visualization

Introduction

High dimensional datasets are difficult to visualize contrary to two or three dimensional data, which can be plotted comparatively easily to demonstrate the inherent structure of the data. To aid visualization of the structure of a dataset, a family of algorithms have been devised in the literature, which are collectively referred as dimensionality reduction algorithms, of which an extensive review can be found in (van der Maaten, Postma, & van den Herik, 2009).

Among these algorithms *t-stochastic neighborhood embedding* (t-SNE) is a novel machine learning technique that has burgeoning applications. t-SNE maps each data point in a given high-dimensional space to a low-dimensional space, typically to a two or three dimensional one, for visualization purposes. The algorithm does a non-linear mapping such that similar points in the high-dimensional space situated nearby each other in the low-dimensional space as well.

In its first stage, the algorithm constructs a probability distribution over pairs of high-dimensional points in such a way that similar points have a high probability of being picked. In the second stage, it constructs the same probabilities between these points in the low-dimensional space. Finally the algorithm minimizes the difference between these probabilities by minimizing Kullback-Leibler divergence between these two distributions (Van der Maaten & Hinton, 2008).

Inherently, the algorithm preserves the manifold that possibly exist in the high-dimensional data and represents this manifold in low-dimensional space. Indeed, this class of dimensionality reduction algorithms is called "manifold learning". In comparison to the more conventional, linear dimensionality reduction techniques such as *principal component analysis* (PCA), which finds a linear mapping with an objective to find a subspace where the projection of each data point lies as close to the original point as possible, manifold learning algorithms preserve the distance between pairs of points. Because of this the manifolds are preserved as well, whereas with PCA, clusters that are far from each other in high-dimensional space might be merged in low dimensional space.

t-SNE also proves to be useful for technology analysts in monitoring target technologies. Technologies such as batteries and storage, which is the target technology in this article, have multiple characteristics that develop over time. The problem facing the analysts is that most modern data sources are unstructured in character. Unstructured data often indicates that the data is of mixed provenance and quality. Furthermore, readily available data is often a mix of actual performance results, and forecasts of potential future results. Even when performance data is available the data is rarely standardized, and therefore contains incomplete and uncertain data.

Table 1. List of technologies in the database.

Compressed Air Energy Storage (CAES)	Nickel–cadmium (NiCd) battery
Edison (NiFe) battery	Nickel-metal hydride (NiMh) battery
Flow batteries	Nickel–zinc (NiZn) battery
Flywheels	Pumped Hydro
Hydrogen storage	Saltwater (sodium-ion) batteries
Lead-acid battery	Sodium-sulfur (NaS) battery
Lithium–air (Li-air) battery	Supercapacitors
Lithium-ion (Li-ion) battery	Superconducting magnetic energy storage
Lithium–sulfur (Li-S) battery	Zinc-air battery

Table Table 1 shows typical sources used in appraising technological development. The data varies by provenance – it is provided through a mix of academic, commercial, government, non-profit and media organizations. Furthermore, the data itself pertains to technologies at different stages of development, and in different modes of deployment or development. An exemplary data source, discussed in the next section, compiles research and development data concerning storage and battery technologies.

Despite the mixed quality of the data sources, such data is useful and should be incorporated into quantitative analyses. In this paper we are primarily concerned with technometric approaches to modelling technology (Coccia, 2005). In particular we are concerned with utilizing such data to produce technological frontiers. Such frontiers are useful for anticipating the future rate of growth, and can be used for developing coordination mechanisms such as technology roadmaps (Phaal, Farrukh, & Probert, 2004).

Evidence and belief need not be mutually incompatible. Bayesian statistical techniques acknowledge that data is often collected in an open, rather than controlled, experimental framework (Gill, 2004). As a result the necessity for belief prevails in the collection of data. There are beliefs concerning the quality of data, the underlying system relationships, and the nature and number of underlying cases to be measured. What is significant then is that prior beliefs concerning the data are acknowledged, that these beliefs actually encompass the true state of the world, and that these beliefs are consistently updated in light of new data. These are requirements which are achievable given the appropriate collection, treatment, and handling of mixed data.

What is required therefore is a technique for handling complexly structured data, for judging cases and similarities, and for managing incomplete data. This paper approaches the problem through the use and extension of *t-stochastic neighborhood embedding* (t-SNE). The technique is used to develop a non-linear manifold of technological performance, and to use this manifold to manage incompleteness in the data. This builds on a long-established technique for handing missing data known as the expectation-maximization procedure (Dempster, Laird, & Rubin, 1977). In the next section, the paper details a database of storage and battery technologies. In the subsequent section, a method is proposed for dealing with this semi-structured data, and in specific, for dealing with uncertain and incomplete technological information.

Data Sources

This work builds upon data collected from Enipedia,[1] a website that collects, organizes and visualizes open data related to energy systems. One of the initiatives on the website has focused on gathering information related to energy storage technologies.

Energy storage is an important topic as many countries are seeking to increase the amount of electricity generation from renewable sources. An issue with renewable energy is that the amount of generation is often variable and can exceed or fall short of the amount that is demanded. If there is an excess of production, then not all of the electricity can be fed into the grid. If there is an undersupply, then power plants relying on fossil-fuels must often be relied on in order to help meet demand. To address this variability, large-scale energy storage could be used to store energy during periods of excess renewable electricity production, and then supply this energy during periods of increased demand.

A key problem is that large-scale energy storage does not currently exist, aside from pumped-storage hydroelectricity plants which can only be built in locations with suitable geography. The development of economically feasible large-scale energy storage technologies will be a major game changer in the energy sector as it can support a larger integration of renewables and decrease reliability on electricity generation from fossil sources.

The research indicated that a number of energy scenarios and simulations fail to include models on energy storage, and lack accurate data on technologies. Also, forecasting is often not included, while battery technologies and costs are rapidly evolving. By these needs, an accessible and open technology database was created, incorporating a total of 18 energy storage technologies and 134 facilities or technology pages with a total of over 1,800 properties. In this database,[2] information on technical maturity, technology readiness level and forecasting is included for a number of technologies.

An overview of sources of technology information on the potential and future demand for energy storage indicates that a number of technologies and solutions focus on applications with small time-scales, such as frequency and voltage control, load shifting, diurnal storage, output smoothing, mobility and reserve grid capacity. Far few technologies and facilities focus on providing seasonal and large-scale grid storage. For a number of these technologies, installations with a lower technology readiness level have been included to provide some numbers on feasibility.

Developing metrics on comparing these technologies was done through an iterative design scheme, incorporating metrics relevant to a range of applications. It was observed that a number of technologies cannot be described fully, as information is missing or the ranges in which information sources report the information are exceptionally wide. Also, the definitions found for some technologies, such as Li-ion, are weaker than those found for other

[1] http://enipedia.tudelft.nl

[2] http://enipedia.tudelft.nl/wiki/Electricity_Storage

technologies. Furthermore, metrics are often made available on a systems level, and information on other levels needs to be translated to this system level.

Table 2. Variable number, name and description

No.	*Variable Name*	*Description*
1	Case	Case number
2	Product	Product name
3	Technology	Technology type
4	Year	Reference year
5	Institutional Data	Indicator whether observation is institutional
6	Technology Readiness Level[3]	Technology maturity level
7	Investment per Unit Power	Investment unit power (EUR/KW)
8	Investment per Unit Energy	Investment cost per unit energy (EUR/KWh)
9	Efficiency	Energy efficiency
10	Cycles	Life span in cycle times
11	Energy Density	Energy density (WH/L)
12	Power Density	Power density (WH/Kg)
13	LCoE[4]	Levelized cost of energy

Method

The chief challenge in managing unstructured data is understanding similarities between technologies. This in turn requires techniques for analysing local structures in high dimensional data. The technique of choice for this is t-stochastic neighborhood embedding (van der Maaten & Hinton, 2008). Finding a manifold which represents the data is useful for developing lower dimensional representations of the data. Such a manifold is inherently non-linear, and by necessity it preserves the local structures in the data at the expense of finding any global structures which might be present. For this analysis we adopt an implementation of the algorithm created in Matlab (van der Maaten, 2007).

The t-SNE technique has previously been used in technometrics. Cunningham and Kwakkel (2014) investigate a case of electric vehicle and hybrid electric vehicle designs and technologies. The case benefitted from the use of a non-linear fitting technique since the designs considered differ substantially in fundaments. As a result different designs highlight fundamentally distinct kinds of engineering trade-offs. The case also demonstrated a potential convergence across multiple technologies. Other patterns of technological evolution on a manifold, in addition to convergence, are identified in the paper.

Other technometric approaches utilize a linear, or quasi-linear technological frontier. Many of these approaches also assume a constant rate of technological change as the frontier advances over time. These alternative approaches are useful for single technologies with well-understood morphologies. Such techniques are also suitable for technologies where there are suitable indicators of performance, outcome, or merit. The techniques are less useful for analyzing broader fields with a heterogeneous base of technology. In such fields different technological trade-offs may be at work, and the pace of technological change may be discontinuous or punctuated. Indeed, the technologies themselves each may be valued for different purposes and outcomes.

[3] http://en.wikipedia.org/wiki/Technology_readiness_level

[4] http://en.wikipedia.org/wiki/Cost_of_electricity_by_source

A desirable method must be suitable for use with diverse data types. Before applying t-SNE to the data set of Table 2, the data is first transformed and normalized. Transforming the data eases a search for locally similar data points. Furthermore, the normalization of the data helps address difficulties associated with variables being measured in different units, potentially highly discrepant in scale. The choice is made to take the logarithm of the data whenever the data is right skewed. Logistic transformation is used to create more normal-like distributions than the actual.

As previously noted, a major challenge in addressing such data sets is the presence of missing data. The principle technique for handling missing data in the statistical literature is known as the expectation-maximization procedure. This powerful technique has been extended to address the estimation of missing model parameters, as well as missing data, and later become a mainstay of machine learning techniques. Modern machine learning procedures are now availed of much faster algorithms than expectation-maximization procedures; nonetheless the technique has had a powerful effect on the field.

The expectation-maximization procedure consists of two steps. In the first, or expectation step, the missing data is replaced with an expected value. Initially the expected value can be set by the mean of the data, or even by replacing the missing data with random values. Then in the maximization step, a model of the data is selected and applied. After an initial modeling step, further estimates of expected values derived from the model can be derived. These expected values become new expected values for additional rounds of the modelling procedure. After repeated cycles of expectation and maximization the estimated values converge, and the full model of the data is derived. The technique has the benefit of replacing missing values with neutral values consistent with an assumed model of the data. The technique therefore makes the best use of available data that is possible, rather than excluding whole variables or cases because they are incomplete.

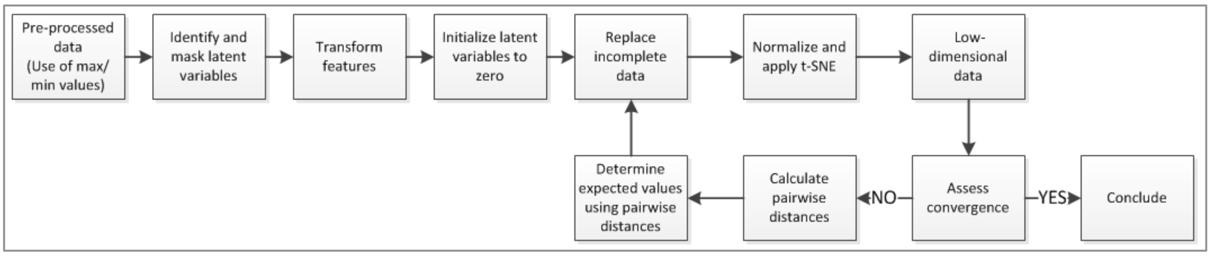
Unstructured data in this domain is not just incomplete, but also uncertain. This is expressed with reported ranges of expected performance data. In order to treat this data, an upper bound and a lower bound on the data is reported, using two distinct model variables. When the data is certain, the upper and lower bound of the variable is identical. In subsequent model runs a constraint is imposed on the expectation maximization procedure – the maximum estimated upper bound on missing data must be greater than the lower bound. When estimated variables do not satisfy this criteria they are either not updated, or both the upper and lower bounds are replaced with averages.

Every point on the manifold estimated by t-SNE is associated with a potential technological design. Thus the t-SNE model is generative – it reports the expected best fit to the data, and also anticipates new cases or designs which have not yet been reported. Nonetheless, technological constraints or other factors may mean that parts of the manifold are not populated with new designs. Interpolation using the manifold can proceed following two directions. A locally linear direction of change can be interpolated from the data given specific examples or cases. Or, a weighted average of surrounding points can be used given their relative proximity on the technological manifold.

Analysis

The following section details a complete procedure for analysis, as depicted in Figure 1Figure. The procedure begins with preprocessing the data. The raw data includes lower and upper bounds for various attributes. Thus, we made a choice to create two different features for each of such variables, e.g., both "Energy density lower bound" and "Energy density upper bound" features.

The next step identifies and masks out the missing data. The process is facilitated by the use of data structures (for instance in Python or Matlab) where the missing data is identified using indicator values. A data matrix therefore contains two layers – the first layer stores the data itself, and the second layer contains a bit matrix for masking. The bit matrix indicates where the data is complete or non-missing, or incomplete and missing.



Pre-processed data (Use of max/ min values)
Identify and mask latent variables
Transform features
Initialize latent variables to zero
Replace incomplete data
Normalize and apply t-SNE
Low-dimensional data
Assess convergence
YES
Conclude
NO
Calculate pairwise distances
Determine expected values using pairwise distances

Figure 1. A Flow Chart of the Analysis Procedure.

Then the features are transformed and normalized to normal-like distributions. The following state initializes the missing variables to zero, which is in effect the mean of the normalized features. In subsequent iterations of the algorithm more refined estimates of the missing data are made. This brings us through the initialization and the first maximization step of the algorithm.

The data is complete, and can now be fitted using the t-SNE algorithm. The major output of the algorithm is a set of coordinates for all the cases – in this example there were 118 points.

Intermediate outputs, such as data coordinates and scatter plots are produced.

Next, convergence of the algorithm is tested by comparing the current imputed high dimensional representation to the high dimensional representation of the previous iteration. Obviously this step is skipped for the first iteration.

If the algorithm has not converged, then pair-wise similarities between the points are evaluated as the next procedure. The purpose of this comparison is to determine the closest peers of any given technology. The basis for this comparison is the Euclidean distance between two points in the three-dimensional space as output from the t-SNE algorithm. The distance is then scaled according to the negative exponential of the squared distance between the two points. The total distance is then re-scaled to sum to 100% percent to create weightings for updating the originally missing variables in the data. The idea here is to calculate the new values for the missing data such that these values are closer to the related data points implied by the low dimensional data. Using pair-wise distances, a new expected set of values is established and finally the high dimensional representation is updated. The model converges when there is negligible differences between the consecutive imputed high dimensional representations.

Results and Visualization

This section discusses some results of the t-SNE analysis, visualizes and interprets some of the results, instead of all, due to space limitations, and displays the technologies according to their respective dates of introduction or their forecasted date of introduction. These colors suggest that the frontier of technological performance is gradually moving outward (to the upper right) over time. This is further illustrated in Figure 3.

Technological development, at least as measured by year of introduction is a somewhat noisy variable. Nonetheless, in Figure 3, we can qualitatively place three frontier lines. The first is dated 10 1985, the second to 2010, and the third to 2035. It seems plausible given the figure that the rate of technological change is higher among battery technologies than it is among storage technologies. This is demonstrated by the comparative “fanning out” of the battery technologies over time.

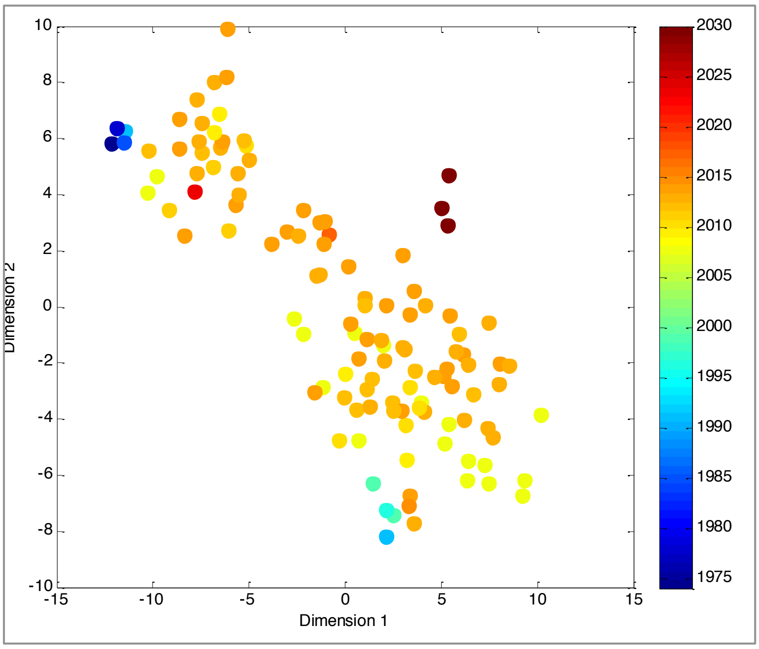


Figure 2. Technologies Positioned by t-SNE and Colored by Date of Introduction

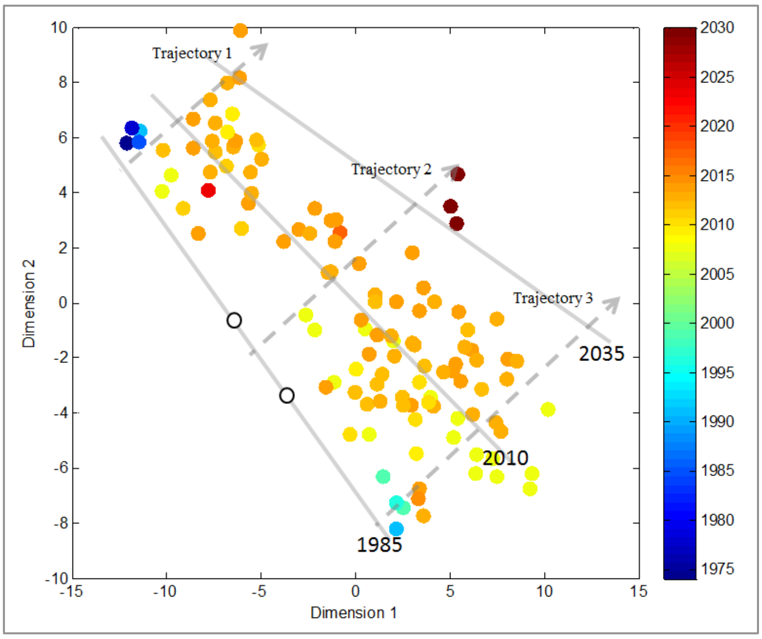


Figure 1. Technological Trajectories

In Figure 3 three technological trajectories are displayed. Changes in technological performance, based on benchmark technologies on or near the trajectory are calibrated. Then the three trajectories are compared with one another to determine whether there are common elements of change across the trajectories.

Figure 4 describes a potential trade-off in the design and selection of battery and storage technologies. In general the trade-off is between the respective cost and advantages of storage technologies versus batteries. Storage technologies are more robust, providing more cycles of operation at a lower levelized cost of energy. This comes at the cost of having a lower energy density, a lower technology readiness level, and a lower efficiency. In contrast battery technologies offer more energy density, are more readily available on the market, and operate at a higher level of efficiency. In consequence, batteries are less robust, operating for fewer cycles, and requires a higher levelized cost of energy to be paid out.

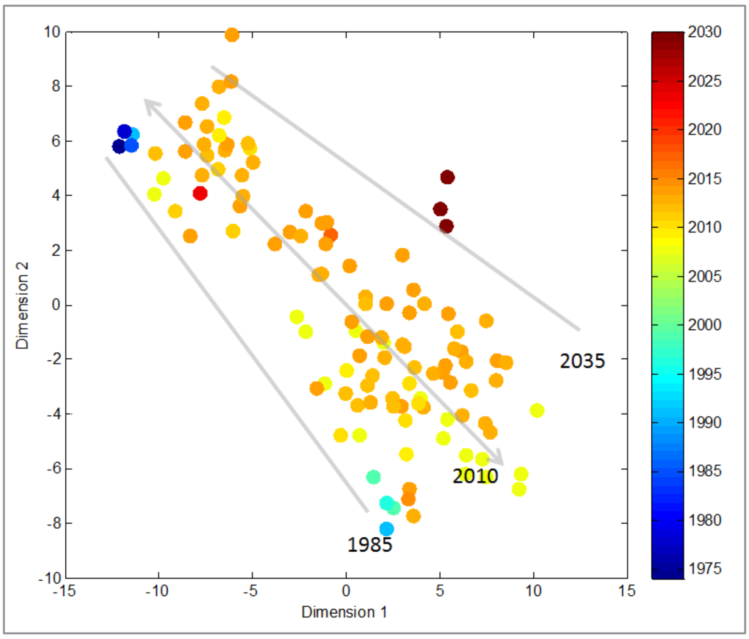


Figure 4. Design Trade-Offs.

There are three design voids on the manifold as shown in Figure 5. These are areas in the space of potential design which have not been explored. One space, design void 1, occurs along the 1985 technological frontier. The space is sparsely explored, although by 2010 a flywheel technology has emerged to occupy the space. The next two voids lie along the 2035 frontier. Because we are not yet on the 2035 frontier, these voids may be unanticipated breakthroughs. Design void 2 is in the space of high performing storage systems, and design void 3 is in the space of high performing batteries. One organization, EASE, anticipates a number of 2030 battery technologies on or beyond this frontier.

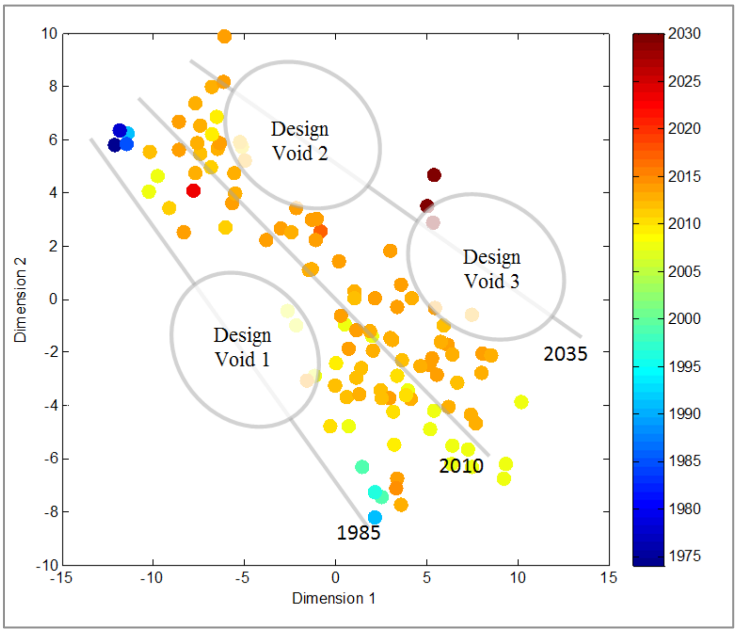


Figure 5. Design Voids.

Table 1. Historical and Emerging Designs.

	Void 1	*Void 2*	*Void 3*
Year	2013	2012	2030
InstitutionalData	0.01	0.79	0.99
TRL	8	6	9
Investment lowerbound	1,093	69	103
Investment upperbound	1,149	131	147
InvestmentEURperKW lowerbound	1,244	729	574
InvestmentEURperKW upperbound	1,262	1549	898
Efficiency lowerbound	0.767	0.709	0.785
Efficiency upperbound	0.849	0.809	0.847
Cycles lowerbound	4,265	11,306	3456
Cycles upperound	4,554	70,551	9804
EnergyDensity lowerbound	40	5	105
EnergyDensity upperbound	60	11	186
Power Density lowerbound	131	82	158
PowerDensity upperbound	220	210	295
LCoE lowerbound	0.149	0.074	0.056
LCoE upperbound	0.506	0.224	0.123

Table 3 provides, by interpolation, the performance characteristics of the technologies in the three voids mentioned previously. The exemplary void 1 technology is most likely a battery. The year of introduction suggests that there have been too few lower technology exemplars, so that the performance here is likely highly overstated. There should likely be a lower power and energy densities, and a lower levelized cost of energy. The closest existing technology is the "Wemag AG Li-Mn storage plant."

The void 2 technology, likely a storage device, should afford dramatically reduced investment and investment per kilowatt hour over previous technologies. The cycle times should be up to an order of magnitude higher than the void 1 exempla. While the power density may not be affected much from its 1985 peer, the energy density is likely to be reduced. The levelized cost of energy may be half of the previous levels of the void 1 technology. The year of introduction is too early, suggesting still higher energy and power densities over those listed. The closest existing technology is an advanced compressed air energy storage device.

The exemplary void 3 technology is most likely a battery. It will require an order of magnitude less unit investment, although the investment in terms of euros per kilowatt may be up to one half of previous levels. Cycle times will be improved, and energy densities may be doubled or even tripled over previous technologies. Power densities will also be somewhat improved. The levelized cost of energy will be three or four times lower than the equivalent technologies from 1985. The technology as anticipated is closest to some of the forecasted lead-acid battery advances for the year 2030.

Conclusions

In this paper, a database of energy storage technologies with various corresponding attributes is examined. The authors described a method to manage incompleteness of the data. The described method synthesizes t-SNE technique, which is a novel dimensionality reduction technique, with long-established expectation-maximization technique. The completed database later used for building a technology frontier that shows the progress of technology in time, discussing the design trade-offs in the technology and finally identifying some design voids in the progress of the technology.

The technique described in this paper can be complementary to wide variety of technometrics or evolutionary technology dynamics approaches which make use of high dimensional technology data.

The technique performs better especially in visualization than other dimensionality reduction applications such as feature selection or feature extraction for two reasons. Firstly, it uses expectation maximization to impute the missing variables, which manages the incomplete data in such a way that the imputed variables have minimal weighting in producing the low dimensional map. Hence, it has least effect on the derivation of the lower dimensional map. Secondly, the t-SNE technique itself is a more suitable approach compared to other dimensionality reduction algorithms such as incumbent Principal Component Analysis (PCA). PCA aims to keep variation in the data and does not care about the pairwise relationships between data points, whereas manifold learning techniques such as t-SNE performs better in keeping similarities.

As a follow up to this work, more applications of this techniques next to the technology trajectories and design voids, as showcased in this paper, are yet to be explored. The promise of this technique is its complementary position in various technometrics analysis, which is yet to be fulfilled.

Furthermore, a methodological study regarding the validation of the technique using controlled experiments on a complete data set is on the research agenda of the authors.

Acknowledgement

Authors thank *Big data roadmap and cross-disciplinarY community for addressing socieTal Externalities (BYTE)* project for funding this research.

References

Coccia, M. (2005). Technometrics: Origins, historical evolution and new directions. *Technological Forecasting and Social Change, 72*(8), 944-979.

Cunningham, S. W. & Kwakkel, J. H. (2014). *Technological frontiers and embeddings: A visualization approach.* Paper presented at the International Conference on Management of Engineering & Technology (PICMET), 2014 Portland, Oregon.

Delft University of Technology. (2014). Enipedia. Retrieved June 20, 2015 from: http://enipedia.tudelft.nl.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B, 39*(1), 1-38.

Gill, J. (2004). *Bayesian Methods: A Social and Behavioral Sciences Approach* (Third ed.). Boca Raton, FL, USA: Chapman & Hall / CRC Press.

Phaal, R., Farrukh, C. J. P., & Probert, D. R. (2004). Technology roadmapping -- A planning framework for evolution and revolution. *Technological Forecasting and Social Change, 71*, 5-26.

van der Maaten, L. (2007). An introduction to dimensionality reduction using MatLab. *Report, 1201*(07-07), 62.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579-2605.

van der Maaten, L. J., Postma, E. O., & van den Herik, H. J. (2009). Dimensionality reduction: A comparative review. *Journal of Machine Learning Research, 10*(1-41), 66-71.