Deep learning geodemographics with autoencoders and geographic convolution

Stefano De Sabbata
School of Geography, Geology and the Env.
University of Leicester
University Road, Leicester, LE1 7RH
United Kingdom
s.desabbata@le.ac.uk

Pengyuan Liu
School of Geography, Geology and the Env.
University of Leicester
University Road, Leicester, LE1 7RH
United Kingdom
pl164@le.ac.uk

Abstract

We present two approaches to creating geodemographic classifications using deep neural networks. Both deep neural networks are based on autoencoders, which allow automating dimensionality reduction before clustering. The second approach also introduces the idea of geographic convolution in neural networks, which aims to mirror in the geographical domain the approach of graphical convolution, which has revolutionised image processes in the past decade. To test our approaches, we created a geodemographic classification based on the United Kingdom Census 2011 for the county of Leicestershire and compared it to the official 2011 Output Area Classification. Our results show that the two deep neural networks are successful in creating classifications which are statistically similar to the official classification and demonstrate high cluster homogeneity.

Keywords: geodemographic, census, machine learning, deep learning, autoencoder, geoconvolution.

1 Introduction

Geodemographic classification is a common approach in quantitative geography to exploring the combined spatial distribution of demographic datasets, such as the decennial census in the United Kingdom (Singeton and Spielman, 2014).

Creating geodemographic classifications is a complex and discretional procedure, which includes the analysis and selection of variables for the subsequent computation of clusters. Gale et al. (2016) developed the 2011 Output Area Classification (2011OAC) for the United Kingdom, starting from an initial set of 167 prospective variables from the United Kingdom Census 2011: 86 were removed, 41 were retained as they are, and 40 were combined, leading to a final set of 60 variables. Gale et al. (2016) finally used the k-means clustering approach to create 8 clusters or supergroups, as well as 26 groups and 76 subgroups.

In this paper, we investigate the potential of deep neural networks for creating geodemographic classifications. In particular, we explore the use of autoencoders for dimensionality reduction as a substitute for variable selection and pre-processing.

Moreover, we present the idea of geographical convolution, which aims to explore the potential contribution of higher-scale patterns in creating geodemographic classifications, mirroring in geocomputation the approach of graphical convolution that has revolutionised image processing.

1.1 Terminology

In deep learning literature, the term "feature" is used to refer to what in GIScience and quantitative geography is commonly referred to as "variable" or "attribute", i.e., numerical values representing in a computer system an attribute or a characteristic of an entity in the real world (e.g., as a table column). The term "feature" is commonly used to refer to the input values, whereas the term "feature representation" (or "feature map") is used to refer to subsequent transformations of the values as the information flows deeper through the network.

In this paper, we adopt this terminology and use the term "case" to refer to all the information available about a single real-world entity (e.g., a row in a table).

2 Related work

Deep learning approaches had a transformative impact in a variety of fields, but these have been a somewhat neglected approach in GIScience and quantitative human geography (Harris et al., 2017). That is partially due to most deep learning approaches focusing on supervised learning, while GIScience has primarily focused on unsupervised approaches, as well as modelling and exploratory tasks. Those include geodemographic classification, which commonly adopts traditional unsupervised machine-learning algorithms, such as k-means. However, unsupervised deep learning approaches such as autoencoders have the potential to revolutionise our approach to tasks involving high-dimensional datasets, such as the development of geodemographic classifications.

Autoencoders (Hinton and Salakhutdinov, 2006) are data-specific algorithms, implemented by artificial neural networks that efficiently learn dense feature representations from features (input data) in an unsupervised manner (Liou et al., 2014). Many variants of the autoencoder approach have been proposed, such as convolutional autoencoders and LSTM autoencoders, which have achieved excellent results in addressing challenging problems in image recognition and natural language processing (e.g., Krizhevsky and Hinton, 2011; Li et al., 2015).

Autoencoders learn to compress highly dimensional data (including a large number of features per case) to a low dimensionality (small number of feature representations) through an encoder, and then to reconstruct highly dimensional data (most commonly, the same large number of features per case) from the encoded feature representations through a decoder, while minimising information loss between the original features and the reconstructed ones. As such, the encoder component of an autoencoder can be used very effectively in dimensionality reduction, as a preliminary step to clustering.

Xie et al. (2016) introduced an unsupervised approach namely Deep Embedding Clustering (DEC) which simultaneously learns data features and cluster assignments using a stacked autoencoder. A similar idea was proposed by Chen et al. (2018), who introduced a deep embedding approach to understanging taxi trip purposes based on trip information augmented with contextual data using a stacked autoencoder, and clustering different trip purposes through k-means, leveraging encoded features from the autoencoder.

Recent years have also witnessed an increasing interest in employing autoencoders in land-cover classification (Zhang et al., 2017), points of interest recommendation (Ma et al., 2018) and quality assessment of building footprints for OpenStreetMap (Xu et al., 2017).

3 Deep learning geodemographics

We developed two deep neural networks, based on the DEC clustering algorithm developed by Xie et al. (2016) using the python library Keras[1] and TensorFlow[2] as backend.

For the experiment here presented, we used the 167 prospective variables from the United Kingdom Census 2011 considered by Gale et al. (2016) as features (input data). To obtain a dataset suitable for repeated testing under different parameters, as well as qualitative understanding the resulting clusters, we limited our geographic scope to the 3054 output areas (each containing about 125 households) in the city of Leicester and the county of Leicestershire (total population about one million people). All variables not already provided as percentages were transformed to percentages based on the provided totals except for area, density, mean and median age, and the variable "Day-to-day activities limited a lot or a little Standardised Illness Ratio". We calculated z-scores from the percentage values, to be used as input features for the autoencoder. We found this to be the most effective solution when used in combination with *tanh* activations.

The first deep neural network (*base*) is based on a relatively simple autoencoder composed of six *Dense* encoding layers and six *Dense* decoding layers. As mentioned above, all the encoding layers use *tanh* activations. The 167 input features are encoded to 128 feature representations by the first layer, then to 64, 32, 16, and finally 8 by the subsequent layers.

The decoding layers also use a *tanh* activation, aiming to rebuild the input starting from the 8 encoded feature representations to 16, 32, 64, 128, and finally 167 feature representations.

The autoencoder model was compiled using an *adam* optimiser, *mse* loss function, and *acc* metric. The encoder was then extracted from the autoencoder model and stacked on top of a K-means clustering layer. This overall model is compiled using an *SGD* optimiser and *kld* loss function.

We devised this final set-up based on the approach proposed by Xie et al. (2016) and further varying the number of layers, the number of feature representations, and the activation approach, to minimise the loss function.

We then developed a second deep neural network (*geoconv*), which is structured in the same fashion as the deep neural network above (same number of layers, feature representations, optimizer, loss function and metric) but aims to implement the idea of geographical convolution, by adding before each *Dense* step a custom *Lambda* layer as described below.

3.1 Geographic convolution

Convolutional neural networks have revolutionised image recognition and demonstrated how it is possible to identify shapes and patterns that go beyond the single pixel by applying smoothing functions to images.

We postulate that a similar approach, namely geographic convolution (geoconvolution), can be used when analysing geographic patterns in data representing area objects.

To implement a geoconvolution, we pre-defined a geographic neighbourhood for each census output area, using the PySAL[3] *Kernel* weights function and a 300-meter bandwidth, which has resulted as the most effective in our experiments. Before each *Dense* layer, a geoconvolution is defined in Keras as a custom *Lambda* layer, which calculates weighted average values for the features (or feature representations) based on the geographic neighbourhood of each case (i.e., census output area), and adds (i.e., concatenates) the weighted averages as an additional set of features (or feature representations) for each case. This procedure is analogous to the convolution procedure used on images, but instead of being applied on a pixel matrix, it is applied to a geographic neighbourhood.

For instance, the first layer takes the 167 features as input. For each case, the geoconvolution layer calculates averages for each one of the 167 input features, using its geographic neighbourhood. The newly computed 167 average values are added to the original 167 values, resulting in a total of 334 values. That is what we define as a geoconvolution step.

A *tanh* activation is used, and the values are used as input for the subsequent *Dense* layer, which maps those 334 values to 128 feature representations. A subsequent geoconvolution layer duplicates those 128 values, resulting in 256 values. The *tanh* activation is used, and a subsequent *Dense* layer maps those 256 values to 64 feature representations, and so on for the subsequent layers, as for the previous deep neural network described above. A similar procedure was implemented for the decoder.

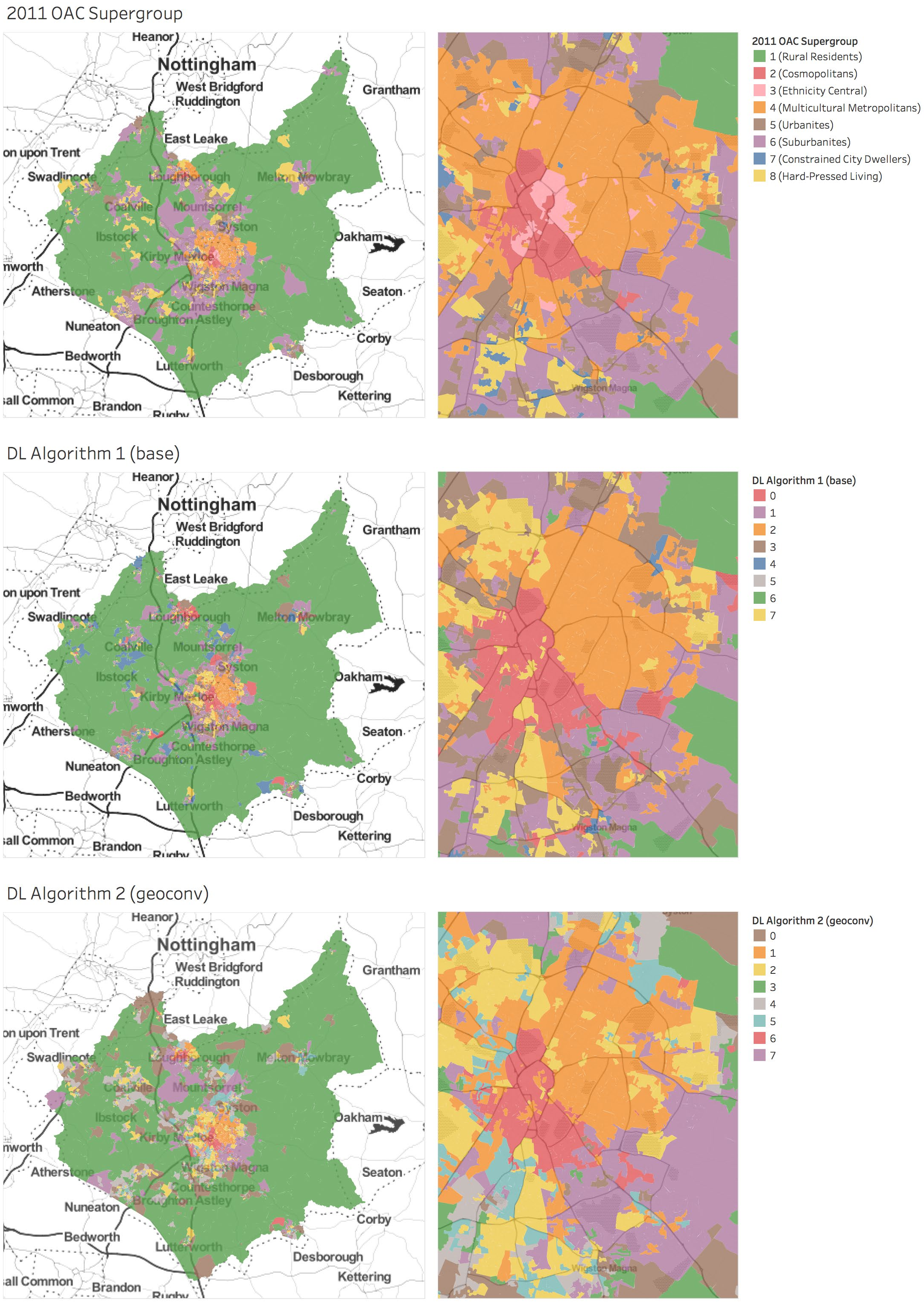
As for the previous deep neural network described above, the encoder was then extracted from the autoencoder model and used in combination with a K-means clustering layer. The overall model was compiled using an *SGD* optimiser and *kld* loss function.

[1] https://keras.io/

[2] https://www.tensorflow.org/

[3] https://pysal.org/

Figure 1: Comparing the 2011OAC and the clusters computed by *base* and *geoconv*, illustrated as maps of the county of Leicestershire (left) and the city of Leicester (right).



Contains data from OpenStreetMap and Office for National Statistics. Map tiles by Stamen Design.

Figure 2: Cross-tabulations illustrating how the 2011OAC has been re-mapped to the clusters computed by *base* and *geoconv* (top) and comparing the clusters computed by *base* and *geoconv* (bottom).

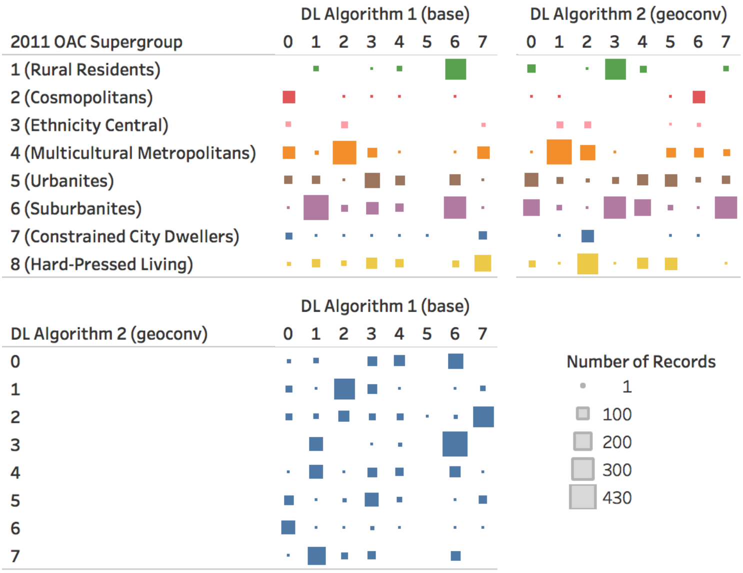
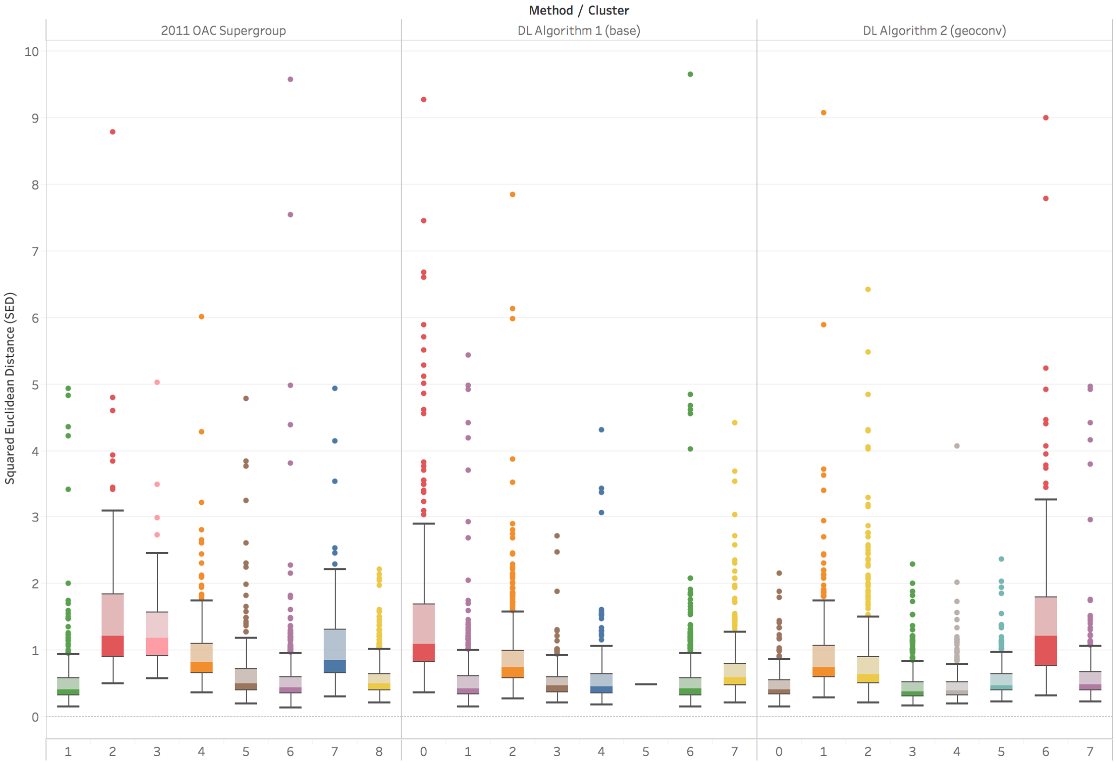


Figure 3: Distribution of Squared Euclidean Distance (SED) scores.



4 Results

As the aim of the experiment here presented was to test the feasibility of a more automated approach to geodemographic classification, we tested our results in comparison with the 2011OAC, as illustrated in Figure 1 and 2.

Figure 1 includes an overview map and a detail of the city of Leicester for each one of the three classifications: the 2011OAC, the classification produced by the first deep neural network (*base*), and the second one using geoconvolutions (*geoconv*). The colours used for the 2011OAC are the same used by Gale et al. (2016), and the colours used for the two classifications here presented have been chosen based on the patterns visible in Figure 2, so to match as far as possible the visual output of the first map.

Figure 2 can be interpreted as a visual representation of the three Chi-Square statistical tests of similarity between the three classifications that we conducted. The results of the tests clearly show that there is a significant association between all three classification: between the 2011OAC and *base*, $X^2(49) = 4667$, $p < 0.001$; between the 2011OAC and *geoconv*, $X^2(49) = 5433$, $p < 0.001$; as well as between *base* and *geoconv*, $X^2(49) = 5678$, $p < 0.001$.

A visual analysis of the figures and maps in Figure 1 and 2 reveals that both *base* and *geoconv* are very effective in recognising the clusters that the 2011OAC interprets as Rural Residents (1) and Cosmopolitans (2), as both have clearly corresponding clusters among those created by the two deep neural networks – cluster 6 and 0 for *base*, 3 and 6 for *geoconv*, respectively. The two proposed deep neural networks also seem fairly capable of recognising the clusters that the 2011OAC interprets as Multicultural Metropolitans (4) and Suburbanites (6), which can be mostly mapped to one or two clusters – 2 and 1 for *base*, 1 and 7 for *geoconv*. It seems quite clear, however, that significant divergences are present. Both deep neural networks seem to map a large portion of output areas from Suburbanites (6) to the same cluster that contains Rural Residents (1) – 6 for *base*, 3 for *geoconv*. Both deep neural networks also seem to cluster most of Constrained City Dwellers (7) along with Hard-Pressed Living (8), as well as some Multicultural Metropolitans (4) – 7 for *base*, 2 for *geoconv*. Ethnicity Central (3) and Urbanities (5) seem the most difficult to recognise for the two deep neural networks here presented, as both create clusters that we couldn't reconduct to any of the classes of the 2011OAC – 5 for *base*, 4 and 5 for *geoconv*.

Finally, we used the squared Euclidean distance (SED) as a measure of cluster homogeneity (Gale et al., 2016). The mean homogeneity score for the 2011OAC in Leicestershire is 0.907, which is slightly higher than the 0.87 score reported by Gale et al. (2016) for the overall UK dataset. The mean homogeneity scores for the two deep neural networks here present are 0.727 for *base* and 0.732 *geoconv*. These values seem to indicate that the clusters created by the deep neural networks are more homogeneous then the ones created for the 2011OAC in Leicestershire, and thus a better representation of the underlying data. Figure 3 provides a more detailed illustration of the SED scores distributions.

Validation is recognised as one of the critical issues in geodemographic research (Singleton and Spielman, 2014), but there are four key limitations which are specific to our evaluation approach as presented above. First and most importantly, the 2011OAC was created using a dataset representing information covering the entire country, whereas the classifications here presented have been created using data covering only Leicestershire. Second, this evaluation assumes that the 2011OAC is a valid classification for Leicestershire, and thus interprets the differences as errors. However, the differences presented above are mostly not drastic, for instance, clustering together output areas that the 2011OAC classified as Urbanities (5) and Suburbanites (6), or Constrained City Dwellers (7) along with Hard-Pressed Living (8). It is possible that at least some of the differences are an improvement on the 2011OAC. Third, the second deep neural network (*geoconv*) uses geoconvolution, which explores patterns in the geographically-local average values that the 2011OAC can't capture. Fourth, we didn't yet attempt an interpretation of the clusters created by *base* and *geoconv*.

5 Discussion

The main contribution of the paper is encapsulated by the results presented in the section above, which clearly illustrates that a largely automated, unsupervised deep neural network can be devised to recreate a geodemographic classification which is statistically similar to the 2011OAC developed by Gale et al. (2016). Furthermore, the clusters computed by the two deep neural networks here presented (*base* and *geoconv*) seem to provide an effective representation of the underlying data, as their average SED scores are relatively low.

The second contribution of the paper is introducing the concept of geoconvolution. We defined geoconvolution as using geographic neighbourhoods to compute convolutions of features representing area objects in deep neural networks. Similar approaches have been shown to be extremely powerful tools for image and language processing (e.g., Krizhevsky and Hinton, 2011; Li et al., 2015). Geoconvolution aims to account for higher-scale patters in the creation of the classification, by looking at the geographically-local average values, whereas common approaches such as k-means are essentially non-spatial. However, the results discussed above don't allow us to identify a clear advantage in using the presented geoconvolution approach (*geoconv*) compared to the base approach (*base*).

The deep neural networks here presented were developed with the objective of recreating the eight classes identified as super-groups in the 2011OAC (Gale et al., 2016). A more general approach, not bounded to that particular number of clusters, or attempting to recreate the more numerous 2011OAC groups or sub-groups, could have led to a different quality of the results. This will be the main focus of our future research.

Finally, the number of possible approaches to implementing the general idea of geoconvolution is vast, and further work is needed to explore this new research avenue fully. While the approaches here presented provide a more automated geodemographic procedure, the number clusters and their interpretation are still largely at the discretion of the practitioner creating a classification, along with the large number of the parameters required to define the deep autoencoders.

References

Chen, C., Liao, C., Xie, X., Wang, Y. and Zhao, J. (2018). Trip2Vec: a deep embedding approach for clustering and profiling taxi trip purposes. Personal and Ubiquitous Computing, pp.1-14.

Gale, C.G., Singleton, A., Bates, A.G. and Longley, P.A. (2016). Creating the 2011 area classification for output areas (2011 OAC). *Journal of Spatial Information Science, 12*, pp.1-27.

Harris, R., O'Sullivan, D., Gahegan, M., Charlton, M., Comber, L., Longley, P., Brunsdon, C., Malleson, N., Heppenstall, A., Singleton, A. and Arribas-Bel, D. (2017). More bark than bytes? Reflections on 21+ years of geocomputation. *Environment and Planning B: Urban Analytics and City Science, 44(4),* pp.598-617.

Hinton, G.E. and Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313(5786)*, pp.504-507.

Krizhevsky, A. and Hinton, G.E. (2011). Using very deep autoencoders for content-based image retrieval. In *ESANN.*

Li, J., Luong, M.T. and Jurafsky, D. (2015). A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint,* arXiv:1506.01057.

Liou, C.Y., Cheng, W.C., Liou, J.W. and Liou, D.R. (2014). Autoencoder for words. *Neurocomputing, 139*, pp.84-96.

Ma, C., Zhang, Y., Wang, Q. and Liu, X. (2018), October. Point-of-Interest Recommendation: Exploiting Self-Attentive Autoencoders with Neighbor-Aware Influence. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management,* (pp. 697-706). ACM.

Singleton, A.D. and Spielman, S.E. (2014). The past, present, and future of geodemographic research in the United States and United Kingdom. *The Professional Geographer, 66(4)*, pp.558-567.

Xie, J., Girshick, R. and Farhadi, A. (2016), June. Unsupervised deep embedding for clustering analysis. *In International conference on machine learning,* pp. 478-487.

Xu, Y., Chen, Z., Xie, Z. and Wu, L. (2017). Quality assessment of building footprint data using a deep autoencoder network. *International Journal of Geographical Information Science, 31(10),* pp.1929-1951.

Zhang, X., Chen, G., Wang, W., Wang, Q. and Dai, F. (2017). Object-based land-cover supervised classification for very-high-resolution UAV images using stacked denoising autoencoders. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10(7),* pp.3373-3385.