sciendo

Journal of Official Statistics, Vol. 35, No. 3, 2019, pp. 625–651, http://dx.doi.org/10.2478/JOS-2019-0026

A Lexical Approach to Estimating Environmental Goods and Services Output in the Construction Sector via Soft Classification of Enterprise Activity Descriptions Using Latent Dirichlet Allocation

Gerard Keogh[1]

The research question addressed here is whether the semantic value implicit in environmental terms in an activity description text string, can be translated into economic value for firms in the construction sector. We address this question using a relatively new applied statistical method called Latent Dirichlet Allocation (LDA). We first identify a satellite register of firms in construction sector that engage in some form of environmental work. From these we construct a vocabulary of meaningful words. Then, for each firm in turn on this satellite register we take its activity description text string and process this string with LDA. This *softly-classifies* the descriptions on the satellite register into just seven environmentally relevant topics. With this seven-topic classification we proceed to extract a statistically meaningful weight of evidence associated with environmental terms in each activity description. This weight is applied to the associated firm's overall output value recorded on our national Business Register to arrive at a supply side estimate of the firm's EGSS value. On this basis we find the EGSS estimate for construction in Ireland in 2013 is about EURO 229m. We contrast this estimate with estimates from other countries obtained by demand side methods and show it compares satisfactorily, thereby enhancing its credibility. Our method also has the advantage that it provides a breakdown of EGSS output by EU environmental classifications (CEPA/CReMA) as these align closely to discovered topics. We stress the success of this application of LDA relies greatly on our small vocabulary which is constructed directly from the satellite register.

Key words: Latent dirichlet allocation (LDA); environmental goods and services (EGSS); satellite register; lexical analysis; supply side estimates.

1. Introduction

Whether it is carbon emissions, increasing global temperatures or the depletion of natural resources such as woodland or water, it is evident that human activity affects the environment. This development has led to an increasing focus on man-made factors that impact the environment and a consequential need to measure and monitor those factors. Interestingly, the natural tendency is to highlight harmful effects such as pollution or increasing global temperatures while efforts that enhance or sustain the environment, such as insulating our homes or producing energy from renewable sources, tend to be given somewhat less prominence. Evidently, from a policy perspective it is important to be able to combine measures of harmful effects with enhancing and sustainable effects, to gain a

[1] Central Statistics Office, Ardee Road, Rathmines, Dublin 6, Ireland. Email: Gerard.Keogh@cso.ie

fuller picture of human impact on the environment. From a statistical point of view EU Regulation 691/2011 (EU-691 2011) on European Environmental Economic Accounts (EEEA) is a framework to build a fuller picture. It provides for the collection of national level data on harmful factors, such as air emissions and material balances that monitor the use of natural resources, as well as mitigating effects such as environmental taxes (e.g., carbon tax) that dis-incentivise harmful means of production.

Regulation 691/2011 (EU-691 2011) also incorporates a module on Environmental Goods and Services Sector (EGSS). This measures the economic value (gross output) of 'eco-industries' (i.e., 'the green economy'). Under this module, member states in the EU are obliged annually from 2017 onward to report data on output value, exports, employment and gross value added in the production of goods and services that mitigate environmental damage, or manage natural resources in a sustainable way. Accordingly, estimating EGSS is now a looming obligation for National Statistics Institutes (NSIs) within the EU. In this article we set out a completely novel supply side approach to estimate EGSS output. Our approach is based on *lexical analysis* of the textual activity description of each firm held on our national Business Register (BR). We use a relatively new applied statistical tool called Latent Dirichlet Allocation (LDA). We apply this to each firm's activity description in turn, with a view to extracting the 'weight of evidence' of the environmental component from the activity description. The resulting weight is multiplied by the firm's total output to estimate the portion of output that is likely to be purely environmental in origin.

Both the OECD manual (OECD 1999) and Eurostat's Practical Guide for Completion of EGSS Accounts (Eurostat 2015) suggest two approaches to estimate the output value of EGSS. First, the demand side approach is based on National Accounts (NA) expenditure aggregates. Often this can be relatively straightforward in that output is largely the NA expenditure aggregate, or some part thereof, for a particular environmental sector; examples include water services, waste water treatment and waste collection and disposal. Second, in contrast to the NA based demand side methodology, the supply side approach, where practicable, typically takes two forms:

a) Using Structural Businesses Survey (SBS) data from primary suppliers of environmental goods and services (e.g., Prodcom), possibly supplemented by a small survey focussed on businesses in specific industry sectors.
b) Conducting a full specialist survey of businesses active in the green economy; an example is the Green Goods and Services Survey (BLS 2011) conducted by the Bureau of Labor Statistics (BLS) in the United States. We note this survey was only run in the years 2010 and 2011 before being cancelled as a result of budget cuts.

Evidently, both demand and supply side approaches have their strengths and weaknesses. Indeed, while the demand side can provide accurate aggregate values, it cannot readily distinguish between different products or services. Accordingly, it can be difficult to identify the purely environmental component of output in this approach. Meanwhile, even though the SBS based supply side approach can differentiate between different products or services, its coverage of environmentally specific products and services can be limited. To overcome this limitation some NSIs will supplement SBS sources with a small survey focussed on specific industry sectors. Of course, a full specialist survey such as the GGSS is likely to yield the most robust estimates of green

output. However, this approach tends to be avoided by smaller NSIs due to attendant costs, and burdens it imposes on respondent businesses.

Estimating EGSS output for a complex sector like construction is particularly problematic, as even well run surveys of the sector are likely to elicit low response levels. Resulting estimates of overall green output are likely to be of poor quality, while CEPA (Classification of Environmental Protection Activities) and CReMA (Classification of Resource Management Activities) breakdowns of EGSS output required under the regulation will be even poorer still. In this situation, a fairly common work-around involves applying appropriate industry specific factors gleaned from experts in the field to existing NA output aggregates (e.g., Statistics Estonia 2015). For example, in the construction sector an appropriate factor might be arrived at by Quantity Surveyors pooling their knowledge of construction costs across a variety of 'standard' construction projects, such as building a typical three-bed home (e.g., RICS 2016). While this approach is sensible it takes little or no account of the specific emphasis of individual firms within the sector. Consequently, without a specific satellite construction register being in place, a firm that is involved in the construction of bridges is treated similarly to one that installs attic insulation, leading to poor estimates and potentially biased breakdowns by CEPA or CReMA.

For a complex sector like construction this seems anomalous and suggests a prerequisite for accurate measurement is the development of an appropriate satellite register, in our case an environmentally specific construction register. If, moreover, we categorise this satellite register by type/class of environmental activity and obtain an expert factor for each class, then we should be able to arrive at a fairly sensible estimate of EGSS output. Clearly, a natural way to determine appropriate classes of environmental activity in construction is to identify meaningful common themes or topics, and use these as a basis for computing an EGSS output estimate. The purpose of this article is to describe how these topics can be learned directly based on a lexical analysis of activity descriptions on our satellite construction register. Furthermore, we show how this knowledge may then be used to arrive at an estimate of EGSS output. We accomplish this using a relatively new applied statistical tool called Latent Dirichlet Allocation (LDA). We show that LDA can learn meaningful environmental topics latent within activity descriptions on our satellite construction register; we note this is a novel application of LDA. Based on the relative importance of these environmental topics within a business's activity description text, we are able to compute an EGSS 'weight of evidence' factor for that business. Importantly, this firm/business level evidence weight reflects the *semantic emphasis* a particular construction business places on those environmental goods and services it supplies. We multiply this weight by the most recent overall supply side output value for that business, as recorded on the BR, to compute an estimate of the value of EGSS output. Summing this across all construction businesses on the satellite register we arrive at an EGSS value for the whole construction sector. We emphasise the 'weight of evidence' that we compute is chiefly a novel by-product of using LDA and therefore renders LDA useful in areas well beyond the field of pure text processing.

The remainder of this article is organised as follows. In Section 2 we discuss the rationale behind our approach, we feel this is necessary because the basis of our approach is quite different to the traditional supply side methodology and justification is therefore needed. Section 3 describes the salient features of LDA's statistical model and Bayesian

inference for this model. This section is a little technical and may be browsed by a reader interested primarily in applications. In Section 4 we describe how we arrive at our satellite construction register and identify a vocabulary of environmentally relevant words based on this register. Section 5 addresses the important issue of model evaluation for topic models. In Section 6 we use the best topic model identified in Section 5 to arrive at an estimate of EGSS output, based on the emphasis of environmentally relevant words in activity descriptions on the satellite register. We compare the estimates with those of other countries and find our overall estimate of EGSS output to be marginally on the low side. We also provide a statistically meaningful further breakdown of our overall estimate by CEPA and CReMA. Section 7 concludes.

2. The Basis of Our Approach

Estimating the weight or portion of a construction business's output that is environmental using LDA is the key novel contribution of this article. The rationale behind this approach is that a firm's description of its activity, as recorded in the activity description itself, stresses the main types of work it carries out. Relative (semantic) weights computed from text analysis of activity descriptions, therefore reflect the relative weight firm's place on different types of work they carry out. Clearly, it is to be expected that the types of work described are also the principle sources of the firm's revenue. Accordingly, it is reasonable to assert that the distribution of relative weights is also reflective of the relative contribution of each type of work (mentioned in the activity description) to the value of economic output. So, for example, a business installing insulation and constructing internal partition walls (i.e., dry walling) has an environmental weight and a pure construction weight. In this article, we show how we can use LDA to compute the relative weight of these two components. Sensibly, we assume these are its principle sources of revenue. The relative weight for insulation computed via LDA reflects the prevalence of insulation in all activity descriptions, in each individual business's particular description, as well as the prevalence of insulation in each topic and across all topics latent in the satellite construction register. Importantly, activity descriptions are classified probabilistically (i.e., a soft classification) according to topic. The amount of total probability that LDA finds in an activity description that is attributable to environmental terms (i.e., words in a description) such as insulation, is a measure of the weight of evidence associated with environmental topics in a particular description and across all descriptions.

Interestingly, the essence of our rationale is that it seeks to mimic the process that an official in a statistics office would apply when forming an impression of the key activities undertaken by a business. Based solely on that business's description of itself and descriptions from similar businesses in that sector, an official would form an impression of the main types of work carried out and their relative importance. Naturally, they would also seek out sensible relevant features (i.e., topics), and use these to arrive at a *refined* sense of the relative importance of the types of work of the business. Further, and in the absence of other sources, if they were required to estimate the distribution of the business's main sources of revenue, with good reason they might adopt weights derived via *refined* relative importance. Of course, the business's self-description may be an activity description as used here, or it may be a description obtained for example from googling the business's website.

We note two particular strengths of our approach that should be highlighted. First, in our implementation of LDA we construct the word vocabulary needed by LDA directly from the satellite construction register itself. This contributes greatly to LDA's success in finding meaningful topics from the activity descriptions on our satellite register. Accordingly, we avoid a common pitfall of using LDA as a *black box* to extract topics from a sea of documents, only to find the topics found have little relevance to real meaningful concepts. Second, as a consequence of finding meaningful topics we are able to map these topics very closely to CEPA and CReMA classifications. Thus, LDA provides us with a statistically meaningful set of relative weights (i.e., distribution) at the activity description level, and therefore at the firm level also, for these classifications. We use these firm level weights to allocate the overall EGSS output estimate according to CEPA and CReMA at the firm level. This is a particularly valuable bonus to adopting our approach.

Significantly, we stress that we do not hold the view that this lexical approach should replace existing demand or supply side methods, but rather provide a complementary method of estimating output from the supply side. Moreover, with this in mind we have programmed the core LDA method in SAS/IML. We feel this may facilitate its wider availability to the official statistics community and in other applied areas such as biostatistics, which often rely on proprietary software systems for statistical analysis. It is also worth mentioning that R has two packages for topic model analysis, one called '*lda'* (Chang and Dai 2015) and the other called '*topicmodels'* (Hornik and Grün 2011). Some of the methodology and analysis conducted here could also be accomplished with these implementations. Equally, proprietary software called MALLET (McCallum 2002) is also available for topic model analysis. As these implementations do not quite fit our needs we have found it expedient to re-work LDA from scratch.

3. Latent Dirichlet Allocation (LDA)

Coding a large data set of natural language textual descriptions (i.e., a corpus of documents) such as business activity, occupation or morbidity is a commonplace job within NSIs. Typically, a coder will *hard code* the description to a single class. Unfortunately, more often than not coding is inexact. In this situation the coder arrives at the appropriate class, via an initial *soft assignment* of two or more classes, based on similarity or relevance judgements and domain specific expert knowledge. The coder then picks the appropriate class from the soft classes based on their relative probabilities or evidence. An important feature of this expert *soft-coding* is its reliance on making an informed choice based on the most sensible combination of relevant themes or topics in the text description. LDA (Blei et al. 2003) is a fully Bayesian procedure that seeks to replicate a first-order approximation to the soft-coding processes of domain experts. Accordingly, it is likely to be of interest to official statisticians and prove beneficial where register development and analysis is needed, as is the case in the realm of EGSS.

The statistical model underpinning LDA relies on a generative model that links a document, labelled by d, in a corpus of D documents, to a set of W unique words in a vocabulary via a latent or hidden set of relevant topics. In LDA topics are typically labelled by the random variable z, the overall number of topics K is constant and assumed *a priori*. The generative model postulates that each document in a corpus is generated

by first picking a multinomial distribution with a K-vector of (topic) parameters $\theta_{jd}(j = 1, \ldots, K)$, where $\sum_{j=1}^{K} \theta_{jd} = 1$. This means the vector of topic parameters θ_{jd} is fixed for the specific document, but varies from document to document in the corpus. For the current document, the K vector of parameters θ_{jd} is generated from a Dirichlet prior distribution with hyperparameter α (which by the way can also be a vector of size K), this, of course, ensures the topic parameters sum to 1. In other words, a prior Dirichlet is used to generate a set of multinomial probabilities θ_{jd} across K topics for the d^{th} document in the corpus. We call the resulting multinomial distribution with parameters $\theta_{jd}(j = 1, \ldots, K)$ generated in this way, the topic multinomial distribution for the document.

Within each topic, LDA's statistical model also specifies a separate multinomial distribution with a vector of parameters φ over all W unique words in the vocabulary. Each individual word in document d is then generated by picking a specific topic $z = j$ from the topic multinomial distribution. This fixes the multinomial distribution with parameter set φ_j for unique words from the vocabulary that occur in topic j, we call this the word multinomial distribution. The individual word in the document is then picked at random from the vocabulary based on the probabilities in this word multinomial distribution. This generative step determines the conditional probability $P(w|z = j) = \varphi_{wj}$ of choosing the word w under the word multinomial distribution for the j^{th} topic. Using the theorem of total probability, we can combine the marginal and conditional probabilities in a mixture model to compute the probability of a specific vocabulary word w as

$$P(w) = \sum_{j=1}^{K} P(w|z_j = j)P(z_j = j) = \sum_{j=1}^{K} \varphi_{wj}\, \theta_{jd} \tag{1}$$

with θ_{jd} the document specific probability (associated with vocabulary word w) in topic j. The full LDA statistical model also posits a Dirichlet prior with hyperparameter β on the (topic specific) word multinomial distribution φ_{wj}, this is used to generate the multinomial word distribution for that specific topic. We mention that the Dirichlet prior distributions are chosen because they are conjugate to the multinomial.

The generative model outlined above may seem somewhat elaborate but in practice it is quite straightforward. Unique words in the vocabulary are assigned probabilistically to a specific topic. Starting with two hyperparameters α and β, we generate a word in a document by first drawing a set of parameters θ_{jd} for topics from a specified $Dirichlet(\alpha)$ distribution; using a Dirichlet prior ensures $\sum_{j=1}^{K} \theta_{jd} = 1$. We then draw (i.e., sample) a specific topic $z = j$ from a multinomial distribution with parameters θ_{jd}. Separately, we draw a set of multinomial parameters φ_{wj} for words in topic j from a $Dirichlet(\beta)$ distribution; once again using a Dirichlet prior ensures $\sum_{w=1}^{W} \varphi_{wj} = 1$. The word is then drawn from the unique vocabulary of words by sampling from the multinomial distribution with parameters φ_{wj}. Repeating this procedure N times generates a document with N words. Further repeating the whole process D times, generates a corpus of D documents where each document is based on K topics.

Consider the following hypothetical example, in the realm of EGSS there might be three topics, *energy saving, renewables and recycling*. For document (i.e., activity description) d, the trinomial topic distribution is generated from a $Dirichlet(\alpha)$ with (probability) parameters $\theta_{d,energy\ saving} = 0.1$, $\theta_{d,renewables} = 0.7$ and $\theta_{d,recycling} = 0.2$. Then, topic $z =$

$2 = renewables$ might be selected based on sampling from this topic distribution. Assuming a 10 word vocabulary, we then generate the multinomial word distribution with (probability) parameters $\varphi_{1,2} = 0.82, \varphi_{2,2} = 0.02, \ldots, \varphi_{10,2} = 0.02$ from a $Dirichlet(\beta)$ for these ten words. Assuming *solar* is the first word in the vocabulary, we then might select it based on these probabilities. This process associates the word *solar* with the topic *renewables* assigned in the d^{th} document. Note, this association of word with topic is purely probabilistic as no other/external information is incorporated. Repeating this procedure N times generates a document with N words from the vocabulary having a trinomial topic distribution and repeating this document process D times generates a corpus based on three topics.

The above process describes how to generate a corpus based on a statistical model. However, in practice, interest centres on using this model as a basis for discovering the set of topics, from an observed corpus of documents and vocabulary of words. This estimation of a set of topics involves learning the matrix parameter sets φ and θ from the words in the corpus of documents. One clever strategy for doing this estimation, introduced by Griffiths and Steyvers (2004), is based on Gibbs sampling (Geman and Geman 1984). Interestingly, this approach avoids sophisticated approximations to difficult integrals of probability distributions that are functions of the parameter sets φ and θ, such as variational Bayes (Blei et al. 2003) or expectation-propagation (Minka and Lafferty 2002). Instead, it seeks to directly evaluate the posterior distribution over the assignments of words to topics $P(z|w)$ and recover the matrices of parameters φ and θ for the corpus of documents by examining this distribution. From Bayes Theorem we can write

$$P(z|w) = \frac{P(w,z)}{P(w)} = \frac{P(w|z)P(z)}{P(w)} \propto P(w|z)P(z) \tag{2}$$

Based on this form, Griffiths and Steyvers (2004) compute separate expressions for $P(w|z)$ and $P(z)$ that are functions of the word-topic counts (n_{wj}) and document-topic counts (n_{jd}) respectively. Using these quantities they derive the full conditional topic distributions required for Gibbs sampling; expressions for the full conditionals, as well as estimates of the (matrix) parameter sets $\hat{\varphi}$ and $\hat{\theta}$ computed from the respective word-topic and document-topic count matrices are given in the Appendix (Section 8). Full details on the derivation of these equations are also given in a number of articles, including Heinrich (2009), Wang (2008), and Carpenter (2010).

Heinrich (2009) also outlines an algorithm for implementing the Gibbs sampler. This too is straightforward as it relies on maintaining matrices for word-topic counts (n_{wj}) and document-topic counts (n_{jd}). The word-topic count matrix (n_{wj}) gives the number of times word w has been assigned to topic j in the vector of assignments z. Meanwhile, the document-topic count matrix (n_{jd}) gives the number of times a word from document d has been assigned to topic j. For the next word in the document, each Gibbs estimation step simply involves decrementing the current count for the topic assignment for that word in both matrices by 1, followed by resampling from the full conditional multinomial topic distribution (i.e., with the current topic excluded) to generate a new topic assignment. The word-topic and document-topic matrices for this word, new topic, and document combination are then incremented by 1. We mention that we have implemented this algorithm in SAS/IML and verified its performance on a novel test problem given in

Griffiths and Steyvers (2004). Interestingly, this test problem comprises 2,000 images (i.e., documents), each being a 5×5 grid of pixels (pixel = word), with the intensity of a pixel specified by an integer and representing the number of times the word occurs in the document. A set of ten topics is constructed; each topic is a 5×5 grid image with a horizontal or vertical white bar set against a black background. Each document is generated by sampling 100 pixels from these topics. The test of our SAS/IML implementation involved generating 500 documents from a vocabulary of 25 words, word 1 to word 25 laid out on 5×5 grid pattern, with these words assigned to topics mirroring those in Griffiths and Steyvers (2004). We ran our implementation for 200 Gibbs iterations and found it recovered the set of ten topics very well indeed, producing results very similar to those reported by Griffiths and Steyvers (2004). We also note that our implementation has two additional refinements; the first allows the Dirichlet prior hyperparameters α and β to be estimated by maximising the joint log-likelihood, Appendix Equations (A1) and (A2), over these hyperparameters via an additional Newton-Raphson step, while the second allows for the Dirichlet topic parameter α to be a vector of length K. We remark however, that in test runs on our EGSS satellite construction register data these refinements only improved on the estimate of the (joint) log-likelihood generated by the core Gibbs estimation routine by a fraction of one percent. In light of this, our analysis proceeds with a scalar topic parameter α, accordingly the $Dirichlet(\alpha)$ and $Dirichlet(\beta)$ distributions are symmetric.

4. The Satellite Register, Document Corpus and Creating the Vocabulary

In our case, the EGSS satellite construction register is a subset of NACE Divisions 41–43 (construction sector) on the CSO's National Business Register (BR). In all, there are over 28,000 entities in the construction sector that describe themselves using approximately 13,500 unique activity descriptions (*Note: after this research was initially completed, a BR coherence project resulted in a substantial increase of approximately 12,000 new businesses in the construction sector being added on to the BR*.). To create the EGSS satellite construction register we have manually scanned each unique activity description and marked it where it included an environmental phrase, such as, *insulation* or *solar* or *heat pump* etc. This process produced a set of 1,077 unique activity descriptions covering 1,228 businesses in the construction sector; our satellite construction register comprises these 1,228 businesses. Meanwhile, we take the set of 1,077 unique activity descriptions we have identified to be our corpus of unique documents (i.e., we simply take each document to be a single unique activity description in this corpus).

In practice, the performance of LDA depends critically on the relevance of the vocabulary. Firstly, we distinguish between words and terms, a word is a unique entry in the vocabulary while a term is the occurrence of a word in a document. Clearly then a word may appear several times as a term in a document. The vocabulary itself is made up of unigram words only, but with some exceptions, such as *heat pump* taken as *heatpump,* while a hyphenation like *Geo-thermal* is taken as *Geothermal*. We follow the practice used in Information Retrieval (IR) and build our vocabulary of relevant words directly from the corpus itself. Initially, each document is first cleaned of punctuation or other non-alphabetic symbols, misspellings corrected and so-called *stop words,* such as THE, IS,

THAT, HE, removed. An initial basic vocabulary of all the unique words is compiled from the terms in the cleaned corpus. Our *cleaned* corpus of 1,077 activity descriptions comprised 5,565 terms corresponding to 830 unique words. We then applied the so-called *tf-idf* scheme (Spärck 1972), a popular scoring method for documents in a corpus, to reduce this further. For each unique word in the vocabulary and each document we compute

$$tf\text{-}idf_{wd} = tf_{wd} \times idf_{w} \tag{3}$$

where tf_{wd} is the term frequency count for word w in document d, and idf_w is the inverse document frequency count, this measures the number of occurrences of each vocabulary word in the corpus (on the log scale). The end result is a word-by-document matrix whose entries are the *tf-idf* values for each vocabulary word in each document in the corpus. The appealing feature of *tf-idf* is that it identifies a set of words that is discriminative for documents in the corpus. Based on the resulting *tf-idf* values, which ranged from about 1.4 to 14, we selected *tf-idf* values of six or higher. This had the effect of removing about 90% of document word instances from the corpus while reducing the vocabulary to 642 words. We further scrutinised the vocabulary words rejected through *tf-idf* analysis and found the 90% cut-off to be too severe as it rejected some words such as *drywall, reclamation, earth, drain, forestry etc.*, which we felt should be in the vocabulary. Accordingly, we decided to scan the remaining 192 unique words and add back some words based on their relevance to construction or environment activity. Of these, we identified seventy unique words that we felt were relevant based on our knowledge of the construction and environmental sectors. Note, while we preferred higher *tf-idf* value words we did not simply select the next highest seventy *tf-idf* values from the 192 unique words. Thus, for example we added back word *UPVC* which has a *tf-idf* value of just 2.845, but is environmentally quite relevant in the fitting of UPVC windows and doors in homes. In any event, this process resulted in a vocabulary comprising 712 unique words that we felt were meaningful construction or environmental words. A full listing of the resulting vocabulary is shown in Appendix Table A2 where we have given a complete list of the *tf-idf* selected vocabulary words and those seventy words added back based on relevance. It is clear from the listing that words added back are relevant and should indeed contribute to improving classification with LDA on our corpus. Moreover, we highlight that the process of compiling the vocabulary was done while assembling a dictionary of environmental terms for EGSS and occurred well in advance of our implementation of LDA. Thus the vocabulary was not selected for LDA so as to specifically fit this corpus, accordingly we stress the results described here are not a consequence of over-fitting using LDA with this vocabulary on our corpus of unique activity descriptions. Interestingly, this vocabulary includes general words like *construction, house* and *system*, as well as more environmentally specific words such as *energy, solar* and *insulate*. This combination of general and specific words in the vocabulary is important, as these combine together probabilistically to generate meaning, and it is topic-specific meaning we are attempting to uncover using LDA. So, having both types of words present in documents will serve to enhance topic learning via LDA. We feed both the corpus of 1,077 unique activity descriptions and the set of 712 unique words in our vocabulary, into our LDA routine with a view to learning or extracting the set of EGSS relevant topics.

5. Identifying the Number of Topics (Model Selection/Evaluation) and Visualising Topics

LDA requires the number of topics K to be given as an input. Accordingly, it makes sense to find an optimum value for K. One method of model selection commonly used to measure performance in IR is to compute the *perplexity* for a subset of held out documents (Heinrich 2009). Roughly speaking, *perplexity* is a cross-validation type measure, found by updating the LDA word-topic and document topic count matrices, via running the Gibbs sampler on an unseen document.

However, *perplexity* has not been adopted widely by statisticians because it does not directly measure the probability or evidence $P(\tilde{d}|K=k) = \prod_{t=1}^{n} P\left(w_{\tilde{d},t}|K=k\right)$ for an unseen held-out document $\tilde{d}$, comprising n terms $w_1, \cdots, w_t, \cdots, w_n$ for words from the vocabulary. Note, generally we use the index j to label topics, but here the topic notation $K=k$ is adopted to distinguish the fact that the number of topics is fixed at k for each computation of the evidence associated with that value of k.

The LDA model assumes documents are independent and words in each document are also independent. Accordingly, from Bayes Equation (2) the evidence is in fact the normalising (probability) constant $P(w)$. Interestingly, Wallach et al. (2009) set out a number of methods to evaluate this quantity for LDA. Their analysis shows a number of methods including the Harmonic Mean Method used by Griffiths and Steyvers (2004) lead to poor estimates of $P(w)$. They offer two credible alternative methods: a Chib-style estimator and their so-called "left-to-right" algorithm. For our purposes we have re-coded their Mathlab Chib-style estimator (see http://people.cs.umass.edu/~wallach/code/etm/) in SAS/IML, with a view to finding an optimal value of K for the corpus of EGSS activity descriptions. Our procedure for finding the optimal K involved running LDA on 90% of the documents in our corpus and holding back 10%. We fixed the number of topics at k and ran LDA on the 90% corpus to get stable estimates of the word-topic (n_{wj}) and document-topic (n_{jd}) count matrices. These were fed into the Chib routine along with the 10% subset of documents held-out, and the evidence probability $P(\tilde{d}|K=k)$ computed for each held-out document. The overall probability for all held-out documents is simply the product of each document's evidence probability, as documents are assumed independent; this independence assumption is valid here, as our documents are activity descriptions from individual businesses that are independent of one another within the construction sector. We simulated this procedure 30 times with a different randomly chosen set of held out documents. This generated 30 estimates for the overall evidence probability. The mean and standard deviation of these 30 estimates is then computed. For accuracy, all computations are done on the log scale, accordingly, we report the Model Log Evidence probability for each setting of k in Figure 1. We mention that to some extent this is a *belt-and-braces* approach, as the resampling in the Chib estimator is designed to give unbiased estimates based on just one simulation.

The plot in Figure 1 shows the results from running the Chib estimation routine. The Mean Log Evidence for each model initially increases as a function of k, reaches a peak at around seven or eight and decreases thereafter. This kind of profile is often seen when varying the dimensionality of a statistical model, with the optimal model being rich enough to fit the information available in the data, yet simple enough to avoid over-fitting

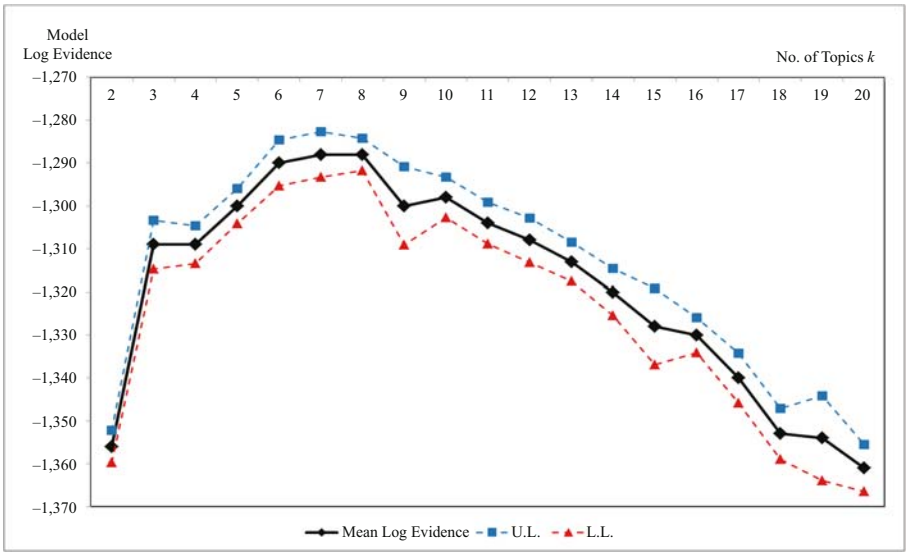

Fig. 1. Model log evidence for topic models.

that data. Typically, in an IR situation, with a corpus of millions of documents and a vocabulary of 10,000 words, finding a small optimal value for k would be unlikely. However, for this dataset the results are very appealing for two reasons. First, as most companies in the construction sector do similar work, we had expected there should only be a small number of topics related to construction within EGSS, and this turns out to be the case. Second, and far more importantly, we expected k to be small because we used a well-defined vocabulary constructed directly from the corpus itself. Accordingly, we expected LDA to find structure based on words having a fairly strong relevance to both EGSS and construction. The plot in Figure 1 also shows the two standard error lower and upper limits, labelled LL and UL respectively, arising from the 30 simulation runs. This band is quite narrow, demonstrating the stability of the Chib estimator and therefore attesting to the quality of the estimated log evidence probability, which is also appealing.

Naturally the value of k found using the Chib procedure depends on the Dirichlet prior hyper-parameters α and β. Each of the 30 runs in our simulation procedure assumed a fixed value k for the number of topics, and α and β initially set equal to 1 and 0.1 respectively. Setting α and β to a fixed constant, is nothing other than a shorthand means of forcing all k parameters in the corresponding Dirichlet distribution to be equal, the resulting distributions are therefore also symmetric. Nonetheless, after running the Gibbs procedure and before running the Chib procedure in each simulation, we also sought optimal values for α and β, given the optimal Gibbs assignments of topics to words in each document. Optimal values for α and β were found by maximising the joint log-likelihood, given in Equations A1 and A2 (Appendix), over these two parameters separately using a Newton-Raphson scheme. The mean values of the resulting estimates of α and β, across each of the 30 simulation runs, is shown in Figure 2 as a function of the number of topics k.

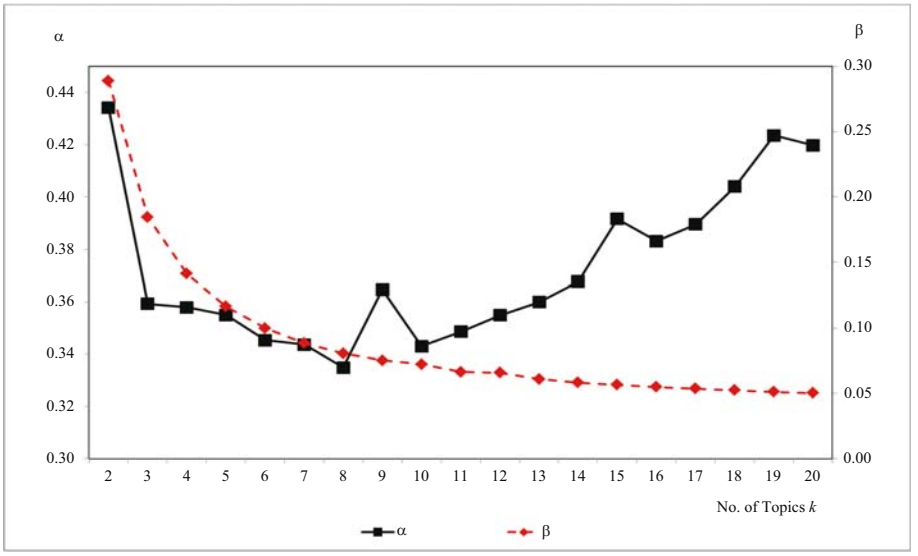

Fig. 2. Mean values for Dirichlet prior hyperparameters α and β for topic models.

It is clear from the plots in Figure 2 that α drops fairly rapidly, reaching a minimum of about 0.33 at about $k = 7$ or 8 topics. Typically, a smaller value for α will favour selecting the same few (i.e., 1 or 2) topic assignments for terms occurring in document d with high probabilities. In practice, this means words in this document can only be assigned to 1 or 2 meaningful topics, and more generally words will therefore tend to cluster strongly according to topic. Thus, as is the case here, when k is relatively low, a small value for α will ensure words cluster into a small number of meaningful topics. Meanwhile, β also drops fairly rapidly with increasing k, but the rate of decent appears to slow significantly at about $k = 7$ or 8 topics with $\beta = 0.07$. This too is appealing, as a small value for β is typical and can be expected to result in a fine-grained decomposition of the corpus into meaningful topics (Griffiths and Steyvers 2004).

Based on this model selection procedure it is clear that sensible settings for the parameters are $K = 7, \alpha = 0.33$ and $\beta = 0.07$, these values are used in all subsequent analysis. The top ten words from the vocabulary associated with this topic model are displayed in Table 1, with the topic titles having been named by us on pragmatic grounds. We *visualise* each topic $k = 1, \ldots, K = 7$, by ranking the words in that topic using their term-score (see Blei and Lafferty 2009)

$$\textit{term-score}_{wk} = \hat{\varphi}_{wk} \times \left(\frac{\hat{\varphi}_{wk}}{\left(\prod_{l=1}^{K} \hat{\varphi}_{wl} \right)^{1/K}} \right) \tag{4}$$

where $\hat{\varphi}_{wk}$ (see Section 8) is the estimated per-topic (vocabulary) word probability. This formula is inspired by the *tf-idf* scheme Equation (3) in that the second term in Equation (4) down-weights words that have high-probability under all topics.

It is clear from Table 1 that topics recovered by running LDA on our activity descriptions are environmentally meaningful. As suggested by the small value found for α

Table 1. Highest ranking term scores for vocabulary words by topic (K = 7).

Windows and doors	Insulation	Agriculture	Pure construction	Water and waste	Alternative energy	Energy saving pollution
WINDOW	INSULATE	BUILDING	TIMBER	WATER	SOLAR	ENERGY
DOOR	ATTIC	CLADDING	HOUSE	CONSTRUCTION	ENERGY	INSTALL
PVC	WALL	CONSTRUCTION	FRAME	WIND	PLUMBING	INSULATE
FITTING	CAVITY	TANK	CONTRACTOR	CIVIL	RENEWABLE	ENVIRONMENT
INSTALL	HOUSE	SLATTED	SITE	TREATMENT	HEAT	AIR
UPVC	BUILDING	SHED	DEMOLITION	HIRE	PUMP	BUILDING ENERGY RATING
GLAZING	ROOF	ROOFING	DRYLINING	SALE	WOOD	ACOUSTIC
ALUMINIUM	CLADDING	FARM	CEILING	INDUSTRY	PELLET	MEMBRANE
REPAIR	FLOOR	AGRICULTURAL	DRAINAGE	TURBINE	GEOTHERMAL	BARRIER
RECYCLING	EXTERNAL	SLURRY	HIRE	WASTE	STOVE	RADON

we see the words cluster nicely within each topic. We also see there is a good degree of discrimination between the topics and a natural degree of overlap, with words like CONSTRUCTION, ENERGY, INSTALL and INSULATE appearing in more than one topic. Meanwhile, words such as REPAIR and RECYCLING are only associated with the "Windows and Doors" topic. This too is quite agreeable, as these words are likely less important in the construction sector than they are in other NACE sectors of EGSS, such as recycling or waste collection and disposal.

Of course, while the results displayed in Table 1 are very appealing, there is the possibility that this topic classification by vocabulary word is to some degree a fluke for this particular value of $K = 7$. Accordingly, the sensitivity of this distribution to the *a priori* set number of topics K is of interest. Ideally, if LDA is stable and the Chib procedure for selecting K is robust, then we should see a topic-word distribution similar to Table 1 for values of K close to 7. To gain some insight into the sensitivity of LDA to the number of topics, we also examined the term score ranking distribution for $K = 6$ and $K = 8$ topics; the distributions are given in the Appendix, Table A1. First, comparing the seven topic distribution in Table 1 with the six topic distribution in Appendix Table A1, we can see a fair degree of similarity. LDA has found three very similar topics; 'Windows and Doors', 'Insulation' and 'Alternative Energy'. However, in this instance LDA has not distinguished words in the area of 'Construction' or 'Water, Waste and Energy Saving' as clearly as it did with $K = 7$ above. This is pleasing as we should see a more course-grained and less relevant set of topics when K is reduced below its postulated optimum value of seven. Second, in the eight topic case LDA seems to work quite well. It has nicely split construction into predominately 'internal' and predominately 'external' construction topics. In the context of the construction sector this seems a pleasing refinement. More importantly, this straightforward sensitivity test shows that LDA is sensitive to the value chosen for K in the best possible way. A small decrease in K from seven to six yields a more course-grained set of topics. However, an increase in K from seven to eight yields a small but appealing alteration to the topics discovered, which remain meaningful from an environmental perspective. We also mention that when we set $K = 20$, we find there are about ten topics that are meaningfully discriminated by LDA, but the other ten are more of a mixed bag. This suggests the Chib procedure is an effective means for determining an appropriate setting for the number of topics K that yields a decomposition of the corpus into meaningful concepts.

6. Using LDA to Estimate EGSS Output Value

The problem we face is simply stated; can we arrive at a meaningful estimate of EGSS value for the construction sector. As remarked in the Introduction, the EGSS Practical Guide (Eurostat 2015) offers no firm method of estimation for EGSS in this sector. Accordingly, EU member states are at liberty to use any credible approach to arrive at a realistic value. With a satellite construction register an ideal solution would be to select a sample and conduct a survey of firms on this register. As noted in the introduction the BLS operated this approach for their GGSS, selecting a sample from a satellite register of about two million environmental businesses in the United States and surveying those selected. An appealing refinement of the GGSS involves using a tailored survey for different sub-sectors, such as the renewable energy or the recycling industry.

Naturally, it would be pleasing to replicate the GGSS, but smaller NSIs typically may not have the resources to operate a specific survey focussed on the environmental sector. Nevertheless, an upper bound for EGSS output in the construction sector is directly available to us. We simply take overall output from the national BR and sum it across all those firms on the satellite construction register. When we do this, we find the resulting output value is EUR 540.8m. Clearly, this is a gross over-estimate of EGSS output, because there are many construction firms where only a portion of their activity is environmental.

Remarkably, if we allow the *data-to-speak-for-itself,* LDA provides a sound approach that enables us to estimate the portion of a construction firm's activity that is genuinely environmental. We turn our focus to the vocabulary and distinguish the subset of words that for all practical purposes are *genuinely environmental* type words, from those that are *essentially construction* sector type words. Examples of *genuinely environmental* words include renewable, solar, insulation and so on, while *essentially construction* words include building, house, construction, and so on. Intriguingly, *essentially construction* words occur frequently in activity descriptions on both the satellite register and the Main BR (NACE Divisions 41-43) covering the whole construction sector. Now, by simply matching the vocabulary with the words occurring on the Main BR, the purely environmental portion of the vocabulary can be tagged and separated from the pure construction portion of the vocabulary. This gives us a vocabulary of *genuinely environmental* type words, upon which we can compute a statistically meaningful weight of evidence favouring environmental activity in each activity description. This weight reflects the semantic emphasis latent in topics that a firm places on the environmental aspects of its own activity description; we refer to it as the semantic weight sem_wt_d.

The Gibbs implementation of LDA maintains a matrix Z of dimension $D \times N$ (i.e., equal in dimension to the document X term matrix) with N being the number of terms in the longest description. This matrix records the most recent topic assignment of the Gibbs Sampler for each term in each document/description. When the Gibbs sampler reaches a steady state, the topic assignments in Z for each term in each document become fixed. Based on these assignments the posterior estimates of word-topic probabilities $\hat{\varphi}_{wj}$ and document-topic probabilities $\hat{\theta}_{jd}$ (see Equations A4 and A5 in the Appendix) are computed. Computing sem_wt_d for d^{th} description proceeds based on these posterior estimates of word-topic and document-topic probabilities. From Equation (1) we can see the probability of each term t, associated with unique vocabulary word w, in each activity description on the satellite register is $\hat{\varphi}_{wj} \times \hat{\theta}_{jd}$. Thus, in steady state, for the t^{th} term in d^{th} description we fix $z = Z_{dt} = j$ for that document-term and compute the corresponding term probability $p_{td} = \hat{\varphi}_{t=w,Z_{dt}=j} \times \hat{\theta}_{Z_{dt}=j,d}$. We define the total term weight to be the sum of these probabilities for all terms matching each vocabulary word w in that description; we label this total term weight for all terms $T(W)_d$, where W is the set of vocabulary words in the description – clearly this sum of probabilities will not in general be equal to one. Similarly, by eliminating the *essentially construction* words from this description, we can identify and retain only those specific term probabilities associated with *genuinely environmental words* in this activity description; we label the resulting total (genuinely environmental) term weight $T(W^*)_d$; by definition this quantity will always be less than or equal to $T(W)_d$ because $W^* \subseteq W$ for all those vocabulary words that appear as terms in

the d^{th} description. We define the evidence favouring environmental strength of meaning in each activity description, as the ratio of the total term weight due to environmental words to that of all words in the description (under the LDA model), this quantity is

$$sem\_wt_d = \frac{P(W^*)_d}{P(W)_d} \tag{5}$$

Multiplying the firm's overall output by this weight gives a statistically meaningful estimate of the output value the construction firm attaches to its environmental activity. Clearly, the stress a firm places on the environmental aspects within its activity description also reflects the economic importance it attaches to these functions. Accordingly, the output estimate derived from directly measuring the relative importance of those environmental aspects via sem_wt_d also has sound economic credibility. We also note the estimated term weight in a description is a linear function of the estimated probabilities p_{td}. Thus, for all words in a given description we have $T(W^*)_d + T(W^\#)_d = T(W)_d$, where $T(W^\#)_d$ is the overall term weight for essentially construction words in that description.

Of course the value of sem_wt_d for activity description depends on the probabilities assigned to *essentially construction* or *genuinely environmental* words in that description. At first sight therefore it would appear that a description with more common *essentially construction* words will have a smaller semantic score than a description with less common *essentially construction* words. This scenario implies that given the same environmental words, the description with more common construction words will be considered less environmental (having a smaller semantic score). This seems anomalous, as the existence of common construction words should not necessarily mean the activity description is less environmental. However, it will be clear from the preceding paragraph that sem_wt_d probabilities are a function of not just of the word, but also the actual topic assigned to that word in the description from the topic assignment matrix Z. Importantly, this varies for the same word across different topics and descriptions. For an essentially 'Construction' topic the scenario outlined above is likely to occur as the probability will be primarily word dependent. But in a topic like 'Alternative Energy' it is far less likely, since topic assignments in the Z matrix will be associated with the respective topic and word in that description. Interestingly, this is quite an appealing feature of LDA as it generates a statistically meaningful score that is dependent on both the type of word and topic assigned to that word at the term level within each activity description.

Implementing this (sem_wt_d) computation, we generated the genuinely environmental vocabulary by removing the set of essentially construction words. The construction words were identified by manually extracting activity descriptions on the satellite corpus that were essentially construction, for example *Carpenter and Builder* and matching these with the 712-word base vocabulary. We found that the base vocabulary was reduced from 712 words to a 249 genuinely environmental word vocabulary. However, in practice we have also found this 249 genuinely environmental word vocabulary turns out to be too restrictive. The reason for this is that certain activity descriptions, such as, "A CIVIL ENGINEERING PLANT HIRE AND DEMOLITION COMPANY" get a zero sem_wt_d value. Interestingly, we included this description on the satellite register as 'demolition' may also incorporate a latent recycling function. Accordingly, to account for this effect

we take the word 'demolition' to be a *latent synonym* for 'recycling' and include it on the genuinely environmental word vocabulary. Identifying and including all these *latent synonyms* our purely environmental word vocabulary increased its size from 249 words to 314 words. Using this extended genuinely environmental word vocabulary we are able to compute a non-zero sem_wt_d for all descriptions on our satellite construction register.

With an effective extended genuinely environmental word vocabulary in place, we computed estimates of the word-topic probabilities $\hat{\varphi}_{wj}$ and document-topic probabilities $\hat{\theta}_{jd}$. In practice, we ran 20 separate LDA simulations on the 1,077 descriptions on the satellite (environmental) register and computed the matrices $\hat{\varphi}_{wj}$ *and* $\hat{\theta}_{jd}$ on each run. Both sets of term weights $T(W)_d$ and $T(W^*)_d$ and the resulting sem_wt_d value, were then computed based on average the values of $\hat{\varphi}_{wj}$ *and* $\hat{\theta}_{jd}$ across the 20 simulation runs. A histogram (and kernel density estimate from Proc SGplot in SAS) of the resulting sem_wt_d values computed for each of the 1,077 activity descriptions is displayed in Figure 3. The plot is skewed to the left with a median value of about 0.41 and lower and upper quartiles of 0.14 and 0.66 respectively. Intriguingly, this tells us that 50% of construction firms on our satellite register, who describe themselves using explicitly environmental words, did so with a degree of environmental *semantic weight or emphasis* below 41%. Meanwhile, only 25% of the firms described themselves with an environmental *semantic weight* of 66% or higher. Recalling that stop-words have been eliminated from our descriptions, the distribution of semantic weight (sem_wt_d) values in Figure 3, suggests that companies in the sector engaged in environmental work, tend to see themselves first as construction companies and second as environmental companies. This shows the NACE coding of these companies, based on their primary activity, into the construction sector tends to be correct, which is a valuable and unforeseen quality assurance by-product of this analysis.

In Figure 4, the Estimated Environmental Output value distribution that results from multiplying each construction firm's output on the 2013 satellite construction register by sem_wt_d is displayed. Encouragingly, as one might expect for output or production value

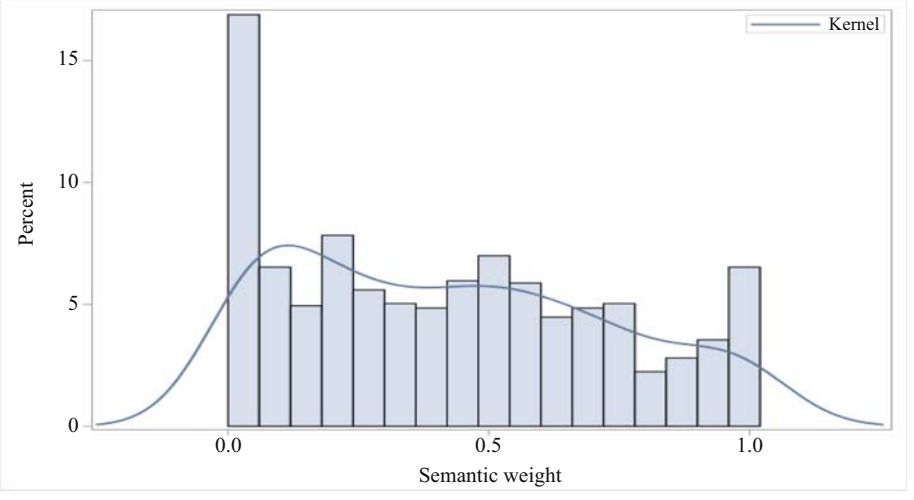

Fig. 3. Distribution of sem_wt_d for activity descriptions on the satellite construction register.

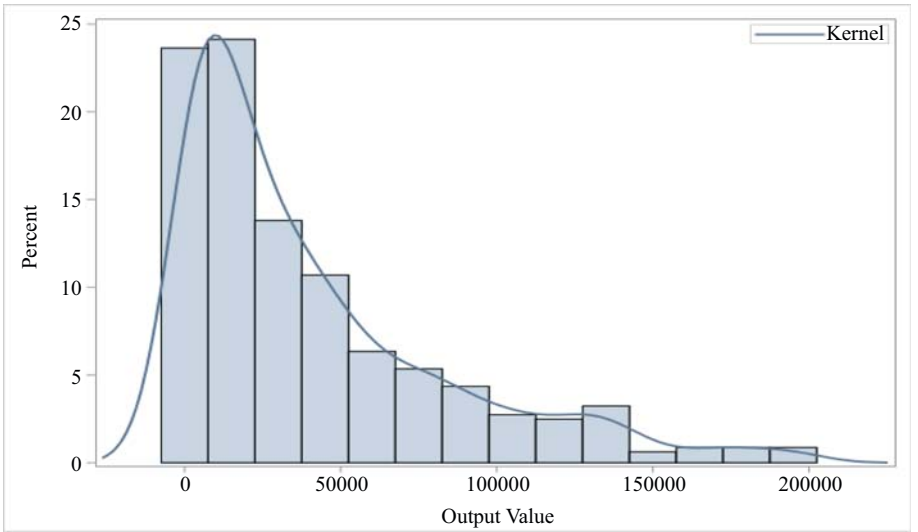

Fig. 4. Distribution of estimated environmental output on the satellite construction register.

data, the distribution is heavily skewed to the left. This is nothing other than a reflection that most construction firms tend to be small companies with a few employees and therefore tend to have a small overall output. The median of this distribution is EUR 36k with the lower and the upper quartiles being EUR 10k and EUR 107k respectively. More importantly, the overall total estimated EGSS output is EUR 229.2m. This value is 42.4% of the overall output value EUR 540.8m for all firms on the satellite construction register. This sem_wt_d based output estimate of EUR 229.2m therefore seems plausible, as many construction companies will typically have quite a mixed bag of activities, several of which will not be environmental. Meanwhile, our current best guess at overall EGSS output for 2013 in Ireland is about EUR 3.4bn, suggesting that the construction component likely accounts for about 6.7% of total EGSS output in Ireland. Interestingly, for international comparison, the 2012 UK EGSS output estimate for construction is 8.3% (ONS 2015) of total environmental output and the 2010 Estonian estimate is 10.5% (Statistics Estonia 2015). If these estimates are accurate, this indicates the sem_wt_d estimate at 6.7% may be somewhat on the low side.

Of course our output estimate is register based and therefore we are able to compute other measures such as output per employee and pay per employee, to further judge the quality of estimated output value. For our satellite register companies these values turn out to be EUR 130.4k and EUR 26.5k respectively, while the corresponding values for general construction companies on the national BR are EUR 119.1k and EUR 24.5k respectively. Considered in this light, our estimates give rise to per employee values that are consistent with the general construction sector in Ireland. Comparing internationally, the UK and Estonia estimates of output per employee are GDP 150.3k and EUR 50.6k respectively, and pleasingly our value of EUR 130.4k comes in close to the middle of these two estimates.

Thus, international evidence based on output per employee and national comparison of pay per employee shows there is a reasonable degree of consistency in our estimated

Table 2. Topic Assignments by CEPA/CReMA.

Topic	CEPA/CReMA Code	CEPA/CReMA class
Windows and doors	13B	Heat/Energy saving and management
Insulation	13B	Heat/Energy saving and management
Agriculture	9, 16	Other environment construction
Pure construction	9, 16	Other environment construction
Water and waste	2, 3	Waste Water management, Waste management
Alternative energy construction	1	Protection of ambient air and climate
Energy saving pollution	13B	Heat/Energy saving and management

output value. Of course the sem_wt_d estimate proposed here is based on emphasis or meaning in text and the degree to which this is causally linked to economic value remains uncertain. Nevertheless, in light of the comparisons made here and the sound statistical methodology (e.g., creating a satellite register and vocabulary construction) underlying the computation of sem_wt_d, it seems reasonable to assume the overall estimate of EUR 229.2m for EGSS in the construction sector in Ireland, based on sem_wt_d, is fundamentally sound.

Remarkably, we can glean more knowledge from our data than just the overall estimate of output value. Specifically, EU Regulation 691/2011 (EU 691,2011) also requires participating member states to provide a breakdown of output value by environmental protection (CEPA) and resource management (CReMA) classifications. Looking at the seven topics identified in Table 1 we can fairly readily associate these with classification headings within CEPA/CReMA as shown in Table 2.

Now, using the estimated document-topic matrix of probabilities $\hat{\theta}_{jd}$ we can allocate the output value of each firm on the satellite register, associated with the d^{th} description, according to these probabilities, giving the CEPA/CReMA value for each firm. Summing these values across all firms we arrive at the EGSS value in the construction sector broken down by CEPA/CReMA, the resulting sector totals are given in Table 3.

The figures in Table 3 show the largest sub-component of EGSS output in the construction sector in Ireland is EUR 98.5m and relates to the area of 'Heat/Energy saving and management'. This covers the provision of insulation and installation of

Table 3. Construction Value by CEPA/CReMA class.

CEPA/CReMA class	Value EUR(m)	Stand Error EUR(m)
Protection of ambient air and climate	32.7	10.7
Waste water management, Waste management	33.9	12.0
Heat/energy saving and management	98.5	20.3
Other environment construction	64.1	16.4
All	229.2	14.7

windows and doors in buildings. Separately we have estimated 'Heat/Energy saving and management' based on aggregate retro-fit insulation grant data (SEAI 2013) and new house construction data (see Department of Environment 2013) and obtained a value of between EUR 100m and EUR 115m. It is pleasing to see that the sem_wt_d estimate of EUR 98.5m computed here is close to the lower end of this interval, adding further to the credibility of our proposed approach. The standard error of the estimated value is also provided based on the 20 simulation runs. The overall standard error of the estimate across all CEPA/CReMA classes comes in at just over 6%, showing the estimated value is quite precise. Also interestingly, this breakdown comes at virtually no additional effort and therefore shows the considerable added value of using LDA to estimate output based on a set of relevant topics. We note that in practice this level of refinement would generally be possible only using a targeted survey such as the GGSS. However, here the time, cost and response burden associated with a specific survey have been avoided.

7. Closing Remarks

The key research problem addressed in this paper is whether and to what extent the semantic value provided in a construction firm's activity description text, informs us about the environmental economic value of the firm. The key assumption underlying this interconnection is, the emphasis a firm places on environmentally related terms in its descriptive text, will also reflect its economic focus and therefore the resulting productive value of the firm. Clearly, in this scenario, the output value of a firm that spends 95% of its time on pure construction work and just 5% of its time on environmental work will be overestimated, if the description comprises several environmental terms and few construction terms. However, this scenario contradicts reality as firms actually do emphasise the activities that are important to them in their activity description. Indeed, like many other NSIs, CSO in its annual BR survey specifically asks each firm to give, *as full a description as possible of its main activities,* and on this basis our underlying assumption seems valid. We also stress that our experience based on purposefully selecting 1,077 environmentally related construction activity descriptions on our satellite register tends to bear this out.

In this article, we used a relatively new applied statistical method called Latent Dirichlet Allocation (LDA) to search for meaning in activity description text strings, in the form of main topics or themes occurring on a satellite construction register. Using the activity descriptions on this register we constructed a vocabulary of 712 unique words needed as input for LDA. We also conducted a model evaluation study and established that our dataset of activity descriptions could be *softly-classified* into just seven environmentally relevant topics. With this seven topic classification we proceeded to extract the weight of evidence associated with environmental terms in each activity description. Based on LDA's estimated word-topic and document-topic probabilities, we proposed a statistically meaningful and environmentally relevant weighting factor. This is based on the ratio of the probability of *genuinely environmental words* in the activity description to the probability of all words; this ratio reflecting the semantic importance of the environmental aspects of the description conditional on the topic.

We applied the resulting evidence weight to the associated firm's overall output value to arrive at an estimate of the EGSS value for each construction firm. The quality of this estimate predicated on the assumption that environmental emphasis placed in the text by that firm, reflects the environmental economic value. On this basis, we arrived at an EGSS estimate for construction in Ireland in 2013 of EUR 229.2m. This accounts for about 41% of the overall output for all firms on the satellite construction register. Comparisons with two other countries, namely the UK and Estonia revealed that the value of our estimate as a proportion of total EGSS value appeared to be on the low side at 6.7%. With this caveat in mind, we viewed the estimated output value of EUR 229.2m for EGSS, arrived at here by analysis of environmental emphasis within activity descriptions, to be fundamentally sound. In addition, we are able to match the topics found by LDA with CEPA/CReMA classes leading to output classified by the latter. This is a valuable extra benefit to using LDA.

It cannot be overemphasised that we have been very purposeful in our use of LDA. Thus, as occurs in many other implementations, we have avoided the pitfall of using LDA as a *black box* to identify latent topics in a corpus of general construction descriptions that then might map to meaningful environmental concepts. Instead, we have pragmatically selected a corpus of environmental activity descriptions and prudently selected a vocabulary based on these descriptions. Thus, from the outset we have done considerable dimension reduction to our dataset before applying LDA. This has put in place the foundations to ensure a meaningful mapping between the topics LDA has discovered and real environmental concepts. Given these operational constraints, our results show that LDA is an impressive tool for identifying meaningful topics. Moreover, we feel this contributes greatly to enhancing the accuracy of our estimates of economic output derived from LDAs document-topic and word-topic probability distributions.

In the literature, LDA and its variants, such as dynamic topic models (Blei and Lafferty 2006), correlated topic models (Blei and Lafferty 2007), tagged or labelled LDA (Ramage et al. 2009) are used solely for text-based corpus analysis. These variants also extend LDA in various ways. By contrast, the analysis conducted here has been undertaken on a relatively small dataset with a small number of topics. Interestingly, we note the approach taken here, where we used a set of purely construction words from a pure construction source, is in essence a form of tagged LDA. Where it is possible to *a priori* tag certain descriptions beforehand with a tag that more precisely identifies economic value with the activity description and/or correlate descriptions, the variants mentioned may give rise to more credible estimates. For this reason and others noted earlier, we stress that the estimate of output arrived at here is not meant to replace estimates arrived at by other (e.g., demand side) means. Ideally, the estimate of EGSS output computed here should complement those others and indeed give a direct breakdown of output according to CEPA/CReMA, as required by the EGSS module in the EU Regulation.

8. Appendix

Using the notation in Section 2, we note from Equation (2) the multinomial distributions over the parameter sets for φ and θ only appear in $P(w|z)$ and $P(z)$ terms respectively. Moreover, as their respective Dirichlet priors are conjugate to these multinomial

distributions, both (matrix) parameter sets φ and θ can be integrated out to give the joint likelihood $P(w, z)$, which is proportional to the product of (see Griffiths & Steyvers 2004)

$$P(w|z) = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta^W)}\right)^K \prod_{j=1}^{K} \frac{\prod_w \Gamma(n_{wj} + \beta)}{\Gamma(N_j + W\beta)} \tag{A1}$$

$$P(z) = \left(\frac{\Gamma(K\alpha)}{\Gamma(\alpha^K)}\right)^D \prod_{d=1}^{D} \frac{\prod_j \Gamma(n_{jd} + \alpha)}{\Gamma(N_d + K\alpha)} \tag{A2}$$

where the entry n_{wj} in the word-topic count matrix (n_{wj}) give the number of times word w has been assigned to topic j in the vector of assignments z, the entry n_{jd} in document-topic count matrix (n_{jd}) gives the number of times a word from document d has been assigned to topic j and Γ is the standard gamma function. Both terms N_j and N_d are the respective topic and document totals of n_{wj} and n_{jd}, while W is the total number of words in the vocabulary. The full conditional topic distributions required for Gibbs sampling are computed from the resulting joint likelihood (see Griffiths and Steyvers 2004). More specifically, denoting the proposed topic to be assigned to term w_t in the d^{th} document by $z_t = (1 \cdots K)$, the full conditional topic distribution associated with this proposed latent assignment z_t is given by

$$P(z_t = j|z_{(-t)}, w) \propto \frac{n_{w_t,j(-t)} + \beta}{N_{j(-t)} + W\beta} \times \frac{n_{j(-t),d} + \alpha}{N_{d,(-t)} + K\alpha} \tag{A3}$$

where $(-t)$ denotes the exclusion of the proposed topic $z_t = j$ for word w_t and $N_{\cdot,(-t)}$ is the total of word-topic and document-topic counts $n_{\cdot,(-t)}$ of the current assignments $z_{(-t)}$ excluding the proposed topic $z_t = j$.

For any single sample we can estimate the word topic and topic document (matrix) parameter sets $\hat{\varphi}$ and $\hat{\theta}$ of probabilities respectively as

$$\hat{\varphi}_{wj} = \frac{n_{wj} + \beta}{N_j + W\beta} \tag{A4}$$

$$\hat{\theta}_{jd} = \frac{n_{jd} + \alpha}{N_d + K\alpha} \tag{A5}$$

The word topic Equation A4 is used to compute the term-score used in Section 4 to visualise words within their topics.

Table A1. Highest ranking term scores for vocabulary words by topic.

K = 6 Topics
Topic

Windows and doors	Insulation	Construction		Water, waste and energy saving	Alternative energy
		Agriculture	Other construction		
WINDOW	HOUSE	BUILDING	INSULATE	WATER	HEATING
DOOR	INSULATE	CLADDING	CLADDING	WIND	ENERGY
PVC	ATTIC	CONSTRUCTION	BUILDING	CONSTRUCTION	SOLAR
FITTING	WALL	TANK	CONTRACTOR	TREATMENT	PLUMBING
INSTALL	TIMBER	SLATTED	ROOFING	TURBINE	INSTALL
GLAZING	FRAME	SHED	HOME	INDUSTRY	RENEWABLE
REPAIR	CAVITY	ROOFING	ROOF	SCHEME	PELLET
UPVC	DRY	FARM	STEEL	BUILDING ENERGY RATING	WOOD
ALUMINIUM	LINING	AGRICULTURAL	SMALL	DISTRIBUTION	GEOTHERMAL
FASCIA	CEILING	SLURRY	ATTIC	RADON	BOILER

Table A1. Continued.

K = 8 Topics
Topic

Windows and doors	Insulation	Agriculture	Construction		Water and waste	Alternative energy	Energy saving and pollution
			Internal	External			
WINDOW	ATTIC	BUILDING	INSULATE	ROOFING	WATER	SOLAR	ENERGY
DOOR	INSULATE	CLADDING	CONTRACTOR	TIMBER	CONSTRUCTION	ENERGY	INSTALL
PVC	WALL	CONSTRUCTION	CLADDING	CLADDING	WIND	PLUMBING	HOME
INSTALL	HOUSE	TANK	DRY	DEMOLITION	CIVIL	RENEWABLE	ENVIRONMENT
FITTING	CAVITY	SLATTED	LINING	CONTRACTOR	TREATMENT	HEAT	AIR
GLAZING	BUILDING	SHED	CEILING	FRAME	HIRE	PUMP	BUILDING ENERGY RATING
UPVC	SMALL	ROOFING	DRYLINING	HOUSE	SALE	WOOD	ACOUSTIC
ALUMINIUM	SALE	FARM	PLASTERING	DRAINAGE	INDUSTRY	PELLET	MEMBRANE
REPAIR	INSTALL	AGRICULTURAL	SLABBING	LAND	TURBINE	GEOTHERMAL	BARRIER
FASCIA	EXTERNAL	SLURRY	DRYWALL	ROOF	WASTE	STOVE	RADON

Table A2. NACE Divisions 41–43 construction vocabulary word list.

Vocabulary words NOT selected by *tf-idf* and added by Author	Vocabulary Word selected by tf-idf scheme ordered by decending *tf-idf* rowwise (Note "S" dropped at word-end)									
EARTH	GRID	CONVERT	GEOTECHNICAL	PHONE	SERVICER	WINDOW	HIGH	COMPANIE	MAINTAIN	UNIT
STONE	DETECTION	CONVERTION	GIO	PHOTO	SHARE	WINDMILL	HOSPITAL	CONVERTING	MANAGEMENT	FITTER
SUSTAINABLE	LEAK	COOLER	GOVERNMENT	PHOTOGRAPHY	SHEEPWOOL	WOODPELLET	IMPORT	DECORATIVE	PARTITION	PRODUCT
LOG	BOREHOLE	CORE	GRAS	PLA	SIGNALLER	WRAPPING	INSTULATION	DELIVERY	PROPERTY	CIVIL
WATERMAIN	SOFTENER	CORRECTION	GROUNDSWORK	PLANTHIRE	SILO	COUNCIL	JOINER	DRILLING	SELLING	SALE
BIOMAS	PUMPING	CORRUGATE	GROWING	PLASTERBOARDIN	SKIM	WELL	JOINTING	DRIVER	SPRAY	SHED
HEATPUMP	VESSEL	COUPLED	HARVESTING	PLASTERER	SLATE	OLD	KERB	DUCT	TILING	CAVITY
GROUNDWORK	WIRING	COVERING	HAULAGE	PLASTIC	SLATTED	PIPEWORK	LED	ERECT	VENTILATION	DOMESTIC
RAIN	SHUTTERING	CRAFT	HEATER	PLU	SOIL	PIT	LIGHT	FOUNDATION	LINE	MAINTENANCE
SEWER	AGRI	CRUSHING	HERITAGE	PLUMBER	SOYA	SEWAGE	MACHINE	GARDEN	RESTORATION	HIRE
DRAIN	AGRIC	CUBICLE	HORSE	PLY	SPRAYER	AUTOMATION	MARKETING	IMPORTATION	DEVELOPER	ROOFING
FIBREGLAS	AUTHORITY	CURATIN	HORTICULTURAL	POLLUTION	SPRING	DRYWALLING	METERING	IMPORTING	ENVIRONMENT	WALL
RECLAMATION	DRYLINER	CUT	HOUSEBUILDING	POLYSTYRENE	STALLER	AIRTIGHT	MILL	LOW	LAND	
MEMBRANE	EXTENTSION	CYLINDER	HVAC	POLYURETHANE	STAT	ALTERATION	OPERATOR	MACHINERY	ACOUSTIC	
RECOVERY	FILL	DAM	IMAGING	POND	STATION	ALTERNATIVE	PARTITIONING	MAIN	CATTLE	
FILTRATION	HYDRO	DAMAGED	IMPROVEMENT	POOL	STEAM	APARTMENT	PARTNER	MECHANICAL	CONSULTANT	
UNDERGROUND	HYGIENE	DAMPROOFING	INFRASTRUCTURE	PORCH	STEELING	APPLICATION	PASSIVE	PIPING	CONTROL	
SEWERAGE	PRESERVATION	DATA	INSPECTION	POTENTIAL	STORM	AREA	PAVING	PREPARE	FIXING	
SLABBING	RUBBLE	DEALING	INST	POURING	STRAW	ASSESSMENT	PLASTERBOARD	PRINCIPAL	FLOORING	
LIGHTING	AEROBOARD	DECOMMISSION	INSTAFIBRE	POWERLINE	STYROFOAM	AUTHORITIE	PROCES	PUBLIC	KITCHEN	
SLAB	AGENT	DEMOLITIAN	INSTRUMENTATION	PRE	SU	BED	PROGRAM	SHOP	LAYING	
SOUND	AGRICULTURE	DENSITY	INTEGRATED	PRIMARY	SUBMERSIBLE	BOARD	PROP	SINGLE	MANUFACTURING	
BARRIER	AI	DESCRIPTION	INTERIOR	PRIOR	SUBSOIL	BONDED	PROPERTIE	SLATING	PAINTING	
FORESTRY	ANALYSI	DESIGNING	INTERPRETATION	PROCESSING	SUDIO	CABIN	PROTECTION	SPACE	PIPE	
BUSINESS ENERGY RATING	ANIMAL	DIG	IRRIGATION	PROD	SUN	CABLE	PROVIDER	SPECIALISE	SAVING	
DRYWALL	APPARTMENT	DISMANTLE	JCB	PROGRAMME	SUPPORT	CHP	PURCHASE	TANKING	SERVICING	
PELLET	ARCHITECTURAL	DOCK	KIT	PROMOTE	SURFACE	CLAY	RADIATOR	TESTING	SUSPENDED	
LANDFILL	ASBESTO	DOM	LAGOON	PUMPED	SURVEYING	CLEAN	RECYCLING	THATCHER	BUSINES	
GROUND	ATTENUATION	DOMESTUC	LANDLORD	PURIFICATION	SW	CLIENT	REFRIDGERATION	TRADE	DISTRIBUTION	
AGRICULTURAL	AUDIO	DRAUGHT	LANDSCAPING	QUOTE	SWEDEN	COAT	RELATED	WATERPROOFING	EXTERNAL	
GEOTHERMAL	AUDIT	DRAW	LARGE	RAFT	SWIMMING	COILLTE	REMOVAL	FILTER	FIRST	
SILAGE	AUTOMATIC	DREDGING	LASER	RAIL	SWITHGEAR	COMMISSIONING	RENTAL	GROUP	METER	
TURBINE	BALING	DRILLING	LEAKAGE	RAINWATER	SYPHONIC	CONSERVATION	REPAIR	PROOFING	PROJECT	
WASTE	BANDED	DRYLINING	LIFT	RAW	SYSTME	CONSERVATORY	RESERVOIR	DIGGER	PROVIDE	
DRYLINING	BASE	DUTIE	LINER	RD	TAPE	HYGIENIC	RETAINING	SPECIALIST	STRUCTURE	
PUMP	BASED	ECOBEAD	LIVE	RECLAIMATION	TAR	CUSTOMER	SAMPLE	SUPPLIE	SUB	
CEILING	BEDROOM	EFFICENT	LOGGING	RECLAMING	TARING	CUTTING	SECURITY	COLD	SUBCONTRACTOR	
ALUMINIUM	BEND	EFFLUENT	MAC	RECOVER	TECHNICALLY	DAIRY	SEPTIC	EXCAVATOR	SUPPLIER	
DRAINAGE	BIN	ELEMENT	MARBLE	REDUCING	TECHNICIAN	DECORATING	SERV	HOT	YARD	
WOOD	BIOFUEL	EMPTYING	MARINE	REED	TECHNOLOGY	DEVICE	SHEET	INTERNAL	CARPENTER	
LINING	BLOCKLAYING	EMULSION	MARKET	REFRIGERATED	TELEPHONE	DISPOSAL	SHIP	LABOUR	SCHEME	
DRY	BLOWN	ENGAGED	MAS	REG	TEO	DISTRIBUTOR	SKIRTING	MATERIAL	ERECTING	
TREATMENT	BODIE	ENGINE	MASTIC	REINFORCED	TEORANTA	DOE	SLOTTED	OPERATION	FARMER	
GLAZING	BOXE	ERRECTION	MATER	REMOVING	THATCHED	DRILL	SMART	PREMISE	HOUSING	
WIND	BREWERY	ESCAVATION	MEASURE	RENOVATING	THERMINAL	DRIVING	SOLID	PRODUCTION	OIL	
SITE	BRICKLAYER	EXPLORATION	METRE	REP	TIGHTNES	DUCTING	SPREADING	REFURBISHMENT	PRIVATE	
FARM	BRIDGE	EXTERIOR	MGE	RESERVOIRE	TILE	DUCTWORK	STAINLES	RETAIL	SECOND	
TIMBER	BURNER	EXTRACT	MGMT	RESLATING	TIPPER	DUMPER	STAIR	SAFETY	SLURRY	
HEATING	BUSINESSE	FACADE	MIDDLE	RESOURCE	TOOL	ECO	STORE	SHEETING	INDUSTRY	
PLUMBING	CALIBRATION	FACILITY	MILKING	RESPRAYING	TOWEE	EFFICIENCY	STRUCTURAL	SOLUTION	EQUIPMENT	
RENEWABLE	CAPPING	FACTOR	MINERAL	RETHATCHING	TRACK	ENERGIE	STRUCTURED	SPECIALISING	CLEARANCE	
FITTING	CARRIED	FARMYARD	MINI	ROADSTONE	TRAILER	ERECTOR	STUD	STORAGE	FABRICATION	
WATER	CCTV	FAST	MITIGATION	ROADWAY	TRANSFORME	ESB	SURVEY	TRACTOR	RATING	
SERVICE	CELING	FEE	MOBILE	ROADWORK	TRUCK	EXTRACTION	TACK	UNDERFLOOR	RESIDENTIAL	
CLADDING	CERTIFY	FELTING	MODULAR	ROCK	TUNNEL	FACTORIE	THATCH	WALLING	ROAD	
SYSTEM	CHAMBER	FEUL	MONITORING	ROCKWOOL	UNDER	FELT	TIMBERFRAME	ASSESSOR	STOVE	
UPVC	CLEANING	FILTERATION	MOVING	ROOD	UNDERTAKING	FIELD	TORCH	EXCAVATION	CONCRETE	
HOUSE	CLEARING	FILTERSCOOLER	NETWORK	ROOFER	UPGRADING	FILM	TRANSMISSION	RENEWAL	ROOF	
SOLAR	COATING	FIREPROOFING	ONLINE	SALVAGE	VISUAL	FIRE	UPGRADE	SLATED	BATHROOM	
CONTRACTOR	COILETE	FIREWOOD	OP	SANITARY	VOLTAIC	FIREPLACE	VALVE	GUTTER	MANUFACTURE	
CONSTRUCTION	COMBINED	FIX	OPERATE	SANITATION	WALK	FOREIGN	VOLTAGE	METAL	PLASTERING	
ENERGY	COMPLEX	FOOD	OPTION	SANITRY	WARDROBE	FOREST	WOODEN	AIR	HOME	
ATTIC	COMPLY	FORMWORK	OVERHEAD	SCANNING	WARE	FRAME	POWER	ASPECT	FLOOR	
BUILDING	COMPONENT	FREE	PARALON	SCRAP	WAREDROBE	FRIENDLY	RADON	BEAD	SMALL	
INSULATE	COMPOSITE	FUEL	PARLOUR	SCREENING	WAREHOUSE	FURNITURE	THATCHING	CRANE	ELECTRICAL	
DEMOLITION	CON	FUME	PATH	SECURING	WATERPROOF	GEO	THERMAL	DWELLING	BOILER	
DOOR	CONDITION	FUMIGATION	PERCULATION	SEI	WATERWAY	GEOLOGICAL	ACTIVITIE	EFFICIENT	FASCIA	
INSTALL	CONDITIONING	FUTURE	PERFORMANCE	SELL	WATERWORK	HANDLING	BIO	FARMING	SOFFIT	
TANK	CONSTRUCTED	GASIFICATION	PERIOD	SEPARATOR	WELDING	HANGING	CENTRAL	FOAM	STEEL	
WINDOW	CONTROLLING	GEOPHYSICAL	PHARMACEUTICAL	SERVER	WHEELIE	HEAVY	CLEAR	GARAGE	HEAT	

9. References

Blei, D., A.Y. Ng, and M. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022. Available at: http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf (accessed May 2016).

Blei, D. and J. Lafferty. 2006. "Dynamic Topic Models." Proceedings of the 23rd International Conference on Machine Learning, 113–120, Pittsburgh, Pennsylvania, U.S.A., June 25 −29, 2006. Doi: https://doi.org/10.1145/1143844.1143859.

Blei, D. and J. Lafferty. 2007. "A Correlated Topic Model of Science." *Annals of Applied Statistics* 1(1): 17–35. Doi: https://doi.org/10.1214/07-AOAS114.

Blei, D. and J. Lafferty. 2009. "Topic Models." Available at: http://www.cs.columbia.edu/~blei/papers/BleiLafferty2009.pdf (accessed April 2016).

BLS. 2011. "Green Goods and Services Survey." Available at: http://www.bls.gov/ggs/, BLS, USA (accessed May 2016).

Carpenter, B. 2010. "Integrating Out Multinomial Parameters in Latent Dirichlet Allocation and Naive Bayes for Collapsed Gibbs Sampling." Available at: https://lingpipe.files.wordpress.com/2010/07/lda3.pdf (accessed May 2016).

Chang, J. and A. Dai. 2015. "'*Package-lda*': Collapsed Gibbs Sampling Methods for Topic Models." Available at: https://cran.r-project.org/web/packages/lda/lda.pdf (accessed May 2016).

Department of Environment. 2013. "Construction Activity Completion Statistics." Available at: http://www.housing.gov.ie/housing/statistics/house-building-and-private-rented/construction-activity-completions, Ireland (accessed April 2016).

Eurostat. 2015. "A Practical Guide for the Compilation of Environmental Goods and Services (EGSS) Accounts." Unit E2, Eurostat, Luxembourg. Doi: https://doi.org/10.2785/688181.

EU-691. 2011. "EU REGULATION (EU) No 691/2011 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 6 July 2011 on European environmental economic accounts." Available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32011R0691 (accessed 2013).

Geman, S. and D. Geman. 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6: 721–741.

Griffiths, T. and M. Steyvers. 2004. "Finding Scientific Topics." *Proceedings of the National Academy of Science*, USA, 101, 5228–5235.

Heinrich, G. 2009. "Parameter Estimation for Text Analysis." Technical Report Fraunhofer IGD, Darmstadt, Germany. Available at: http://www.arbylon.net/publications/text-est2.pdf (accessed April 2016).

Hornik, K. and B. Grün. 2011. "topicmodels: An R Package for Fitting Topic Models." *Journal of Statistical Software* 40(13): 1–30. Doi: https://doi.org/10.18637/jss.v040.i13.

McCallum, A. 2002. "MALLET: A Machine Learning for Language Toolkit." Available at: http://mallet.cs.umass.edu (access May 2016).

Minka, T. and J. Lafferty. 2002. "Expectation-propagation for the Generative Aspect Model." Proceedings of the Eighteenth Conference on Uncertainty in Artificial

Intelligence, 352–359, Alberta, Canada, August 1–4, 2002. Available at: https://dl.acm.org/citation.cfm?id=2073918 (accessed April 2016).

OECD. 1999. "THE ENVIRONMENTAL GOODS AND SERVICES INDUSTRY Manual for Data Collection and Analysis." OECD, Paris. Available at: https://unstats.un.org/unsd/envaccounting/ceea/archive/EPEA/EnvIndustry_Manual_for_data_collection.PDF (accessed May 2016).

ONS. 2015. "UK Environmental Goods and Services Sector (EGSS): 2010–2012." Available at: http://www.ons.gov.uk/economy/environmentalaccounts/bulletins/ukenvironmentalaccounts/2015-04-15 (accessed October 2014).

Ramage, D., D. Hall, R. Nallapati, and C.D. Manning. 2009. "Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora." Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 248–256, Singapore, August 6–7, 2009. Available at: https://www.aclweb.org/anthology/D09-1026 (accessed May 2016).

RICS. 2016. "The Real Cost of New House Delivery, Royal Institute of Charter Surveyors." Dublin, Ireland. Available at: https://www.scsi.ie/documents/get_lob?id=885&field=file (accessed May 2016).

SEAI. 2013. "Sustainable Energy Authority of Ireland – Annual Report 2013." Available at: https://www.seai.ie/Publications/SEAI_Corporate_Publications_/Annual_Reports/SEAI-Annual-Report-2013.pdf (accessed May 2016).

Spärck Jones, K. 1972. "A Statistical Interpretation of Term Specificity and its Application in Retrieval." *Journal of Documentation* 28: 11–21. Doi: https://doi.org/10.1.1.115.8343.

Statistics Estonia. 2015. "Development of the Methodology for the Compilation of Statistics of the Environmental Goods and Services Sector (EGSS) in Estonia." Statistics Estonia.

Wallach, H.M., I. Murray, R. Salakhutdinov, and D. Mimno. 2009. "Evaluation Methods for Topic Models." Proceedings of the 26-th International Conference on Machine Learning", 1105–1112, Montreal, Canada, June 14–18, 2009.

Wang, Y. 2008. "Distributed Gibbs Sampling of Latent Topic Models: The Gritty Details." Available at: https://cxwangyi.files.wordpress.com/2012/01/llt.pdf (accessed May 2016).

Received July 2016
Revised September 2017
Accepted December 2017